# ((( ))) VERSANT™

# Versant™ Spanish Test

Test Description and Validation Summary

# Table of Contents

# 1. Introduction

The Versant™ Spanish Test provides an accurate assessment of how well a person speaks Spanish and understands spoken Spanish. The Versant Spanish Test is intended for adults and students over the age of 18 and takes about 15 minutes to complete. Because the Versant Spanish Test is delivered automatically on a computer or over the telephone by the Versant computerized testing system, the test can be taken at any time, from any location with a land-line phone or a computer with appropriate software. The computerized scoring allows for immediate and objective results that are highly reliable and that correlate with other established measures of spoken Spanish performance.

Academic institutions, corporations and government agencies throughout the world will find the Versant Spanish Test to be useful in evaluating the Spanish listening and speaking ability of students, staff, or applicants.

# 2. Test Description

During the test, the system presents a series of spoken prompts in Spanish at a conversational pace and elicits oral responses in Spanish. The voices that deliver the prompts are from native speakers of several different countries, providing a range of native accents and speaking styles.

The Versant Spanish Test has seven sections: Reading, Repeats, Opposites, Short Answer Questions, Sentence Builds, Story Retelling, and Open Questions. All items in the first six sections elicit responses that are analyzed automatically. These item types provide multiple, independent measures that underlie facility with spoken Spanish, including phonological fluency, sentence comprehension, vocabulary, and pronunciation of rhythmic and segmental units.  Because more than one task type contributes to each subscore, the use of multiple item types maximizes score reliability.

The Versant testing system analyzes the test taker's responses and posts scores online usually within minutes of the completed call. Test administrators and score users can view and print out test results from a password-protected website.

The Versant Spanish Test score report is comprised of an Overall score and four diagnostic subscores: Sentence Mastery, Vocabulary, Fluency, and Pronunciation. Together, these scores describe the test taker's facility in spoken Spanish – that is, the person's ability to understand spoken Spanish on everyday topics and to respond appropriately at a native-like conversational pace in intelligible Spanish.

## 2.1 Test Administration

The Versant Spanish Test generally takes 13 to 17 minutes to complete. Tests can be administered over the telephone or on a computer. Test takers are also welcome to take a practice test if they wish to familiarize themselves with the test prior to the actual test administration.

### 2.1.1 Telephone Administration

Telephone administration is supported by a test paper. The test paper is a single sheet of paper with material printed on both sides. The first side contains general instructions and an introduction to the test procedures (see Appendix). These instructions are the same for all test takers. On the second side is the individual test form, which contains the phone number to call, the Test Identification Number, the spoken instructions written verbatim, item examples, and the printed sentences for Part A: Reading. The test form is unique for each test taker.

It is recommended that the test administrator to give the test paper to the test taker at least five minutes before starting the Versant Spanish Test. The test taker then has the opportunity to read both sides of the test paper and ask questions before the test begins. The administrator should answer any procedural or content questions that the test taker may have.

When the test taker calls into the Versant testing system, the system will ask the test taker to use the telephone keypad to enter the Test Identification Number on the test paper. This identification number keeps the test taker's information secure.

An examiner voice speaks all the instructions for the test. The spoken instructions for each section are also printed verbatim on the test paper to help ensure that test takers understand the directions. Test takers interact with the test system by responding in Spanish as they go through all seven parts of the test as prompted, until they complete the test and hang up the telephone.

### 2.1.2 Computer Administration

For computer administration, the computer must have an Internet connection and the Versant Computer Delivered Test (CDT) software available. The test taker is fitted with a microphone headset. The system allows the test taker 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. Test takers interact with the test system in Spanish, speaking their responses into the microphone. When the test is finished, the test taker clicks a button labeled, "END TEST".

## 2.2 Test Format

Instructions for the test are presented by an examiner voice over the testing system. They are also printed verbatim on a test paper for telephone administration and on the computer screen for computer administration. Test items themselves are presented in various native-speaker voices that are distinct from the examiner voice.

The following subsections provide brief descriptions of the task types and the abilities required to respond to the items in each of the seven parts of the Versant Spanish Test.

## Part A: Reading

In this task, test takers read printed, numbered sentences, one at a time, in the order requested. For telephone administration, the printed text is available from a test paper, which is given to the test taker before the start of the test. For computer administration, the text is displayed on the computer screen. Reading items are grouped into sets of four sequentially coherent sentences as in the example below.

Example:

> 1. Julio había recibido de regalo una hermosa bicicleta último modelo.
> 2. Iba a la universidad y al trabajo en su nueva bicicleta todos los días.
> 3. Un día, al no encontrarla, creyó que se la habían robado.
> 4. Al día siguiente volvió a aparecer: había sido una broma de sus amigos.

Presenting the sentences in a group helps the test taker disambiguate words in context and helps suggest how each individual sentence should be read aloud. The test paper or computer screen presents two sets of four sentences and asks the test taker to read six of these sentences in a random order. The test system tells the test taker which of the numbered sentences to read aloud. After the end of one sentence, the system prompts the test taker to read another sentence from the list.

The sentences are relatively simple in structure and vocabulary, so they can be read easily and fluently by literate speakers of Spanish. Even for test takers with little facility in spoken Spanish but with some reading skills, this task provides samples of pronunciation and reading fluency. The reading task starts the test because, for many test takers, reading aloud presents a familiar task and is a comfortable introduction to the interactive mode of the test as a whole.

## Part B: Repeats

In this task, test takers hear sentences and repeat them verbatim. The sentences are presented to the test taker in approximate order of increasing difficulty. Sentences range in length from about 3 words to 15 words. The audio item prompts are presented by several different voices and are read in a colloquial manner.

Examples:

> ¿Cómo te llamas?
> El joven camina por la calle.
> Le gusta cantar canciones románticas.

To repeat a sentence longer than about seven syllables, the test taker has to recognize the words as produced in a continuous stream of speech (Miller & Isard, 1963). Highly proficient speakers of Spanish can generally repeat sentences that contain many more than seven syllables because these speakers are very familiar with Spanish words, phrase structures, and other common syntactic forms. If a person habitually processes five-word phrases as a unit (e.g., "el gato negro muy gordo" / "the very fat black cat"), 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 less familiar with Spanish sentence structure.

Because the Repeat items require test takers to organize speech into linguistic units, it tests their sentence mastery. In addition, the task has them repeat back full sentences (as opposed to just words and phrases), and therefore, it also offers a sample of the test taker's pronunciation and fluency in spoken Spanish.

## Part C: Opposites

For the Opposites task, test takers hear a Spanish word and are asked to respond with a word that has the opposite meaning. The items are presented in a stratified random order so that the item difficulty generally increases over the sequence of items presented.

Examples:

> alto
>
> subir
>
> blanco

The opposites used in the Versant Spanish Test are quite simple, but they require a mastery of basic Spanish vocabulary in spoken form.

## Part D: Short Answer Questions

In this task, test takers listen to spoken questions in Spanish and answer each of these questions with a single word or short phrase. The questions generally include at least three or four content words embedded in some particular Spanish interrogative 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 Spanish or Latin American 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 Spanish and an adult who has never lived in a Spanish-speaking country.

Examples:

> ¿Cuántas patas tiene un perro?
>
> En una hora, ¿cuántos minutos hay?
>
> Si estuviera enfermo, ¿a quién iría a ver: a un médico o a un vendedor?

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

## Part E: Sentence Builds

For the Sentence Builds task, test takers hear three short phrases. The phrases are presented in a random order (excluding the original word order), and the test taker is asked to rearrange them into a sentence.

Examples:

te  /  María  /  ama

de la mesa  /  el plato  /  recogió

y lo vimos  /  nos asomamos  /  marcharse

For this task, the test taker has to 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 test taker demonstrates facility in spoken Spanish. This measurement is quite distinct from testing memory span (see Test Construct section below).

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

## Part F: Story Retelling

In this task, test takers listen to a story and then are asked to describe what happened in their own words. The test taker is encouraged to tell as much of the story as they can, including the situation, characters, actions and ending. The stories consist of two 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 or event followed by a possible reaction or implicit sequence of events.

Example:

Tres niñas caminaban a la orilla de un arroyo cuando vieron a un pajarito con las patitas enterradas en el barro. Una de las niñas se acercó para ayudarlo, pero el pajarito se fue volando, y la niña terminó con sus pies llenos de barro.

Story retelling responses capture the test taker's listening comprehension ability and measure the test taker's vocabulary range and accuracy in an extended response. The responses also provide samples of spontaneous speech.

## Part G: Open Questions

In the last Versant Spanish Test task, the system presents a spoken prompt in Spanish asking for an opinion, and the test taker provides an answer, with an explanation, in Spanish. The questions deal either with family life or with the test taker's preferences and choices.

Examples:

> ¿Qué opina usted de la contaminación ambiental?
>
> ¿Prefiere usted vivir en la ciudad o en el campo?  Por favor explique su elección.
>
> En su opinión,  ¿es importante que los integrantes de una familia se reúnan a comer juntos alrededor de la mesa, o es mejor que cada uno coma cuando le sea más conveniente?  Por favor explique su respuesta.

This task is used to collect a spontaneous speech sample. The test taker's responses are not scored automatically (as they are in Parts A-F), but rather are available for human review by authorized listeners. Test administrators might choose to have the sample analyzed and scored according to other measures of oral proficiency such as the Common European Framework or CEFR (Council of Europe, 2001).

## 2.3 Number of Items

In the administration of the Versant Spanish Test, the Versant testing system presents a series of discrete items. In total, 60 items are presented to each test taker in seven separate sections. The 60 items are drawn at random from a larger item pool. For example, each test taker is presented with eight Opposite items from among those items available in the pool, but most items will be different from one test administration to the next. Proprietary algorithms are used by the Versant testing system to select from the item pool – the algorithms take into consideration, among other things, the item's level of difficulty and the order of presentation. Table 1 shows the number of items presented in each section.

Table 1. Number of items presented per task.

| Task | Presented |
|---|---|
| A. Readings | 6 (of 8 printed) |
| B. Repeats | 16 |
| C. Opposites | 8 |
| D. Short Answer Questions | 16 |
| E. Sentence Builds | 8 |
| F. Story Retelling | 2 |
| G. Open Questions | 2 |
| **Total** | **58** |

## 2.4 Test Construct

The Versant Spanish Test measures facility in spoken Spanish – that is, the ability to understand spoken Spanish on everyday topics and to respond appropriately at a native-like conversational pace in intelligible Spanish. There are many basic 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).
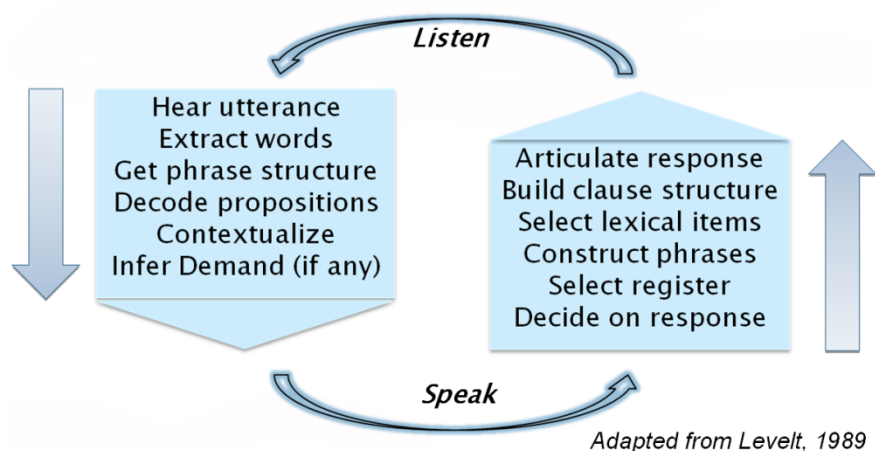


Figure 1. Conversational processing components in listening and speaking.

Core language component processes, such as lexical access and syntactic encoding, can take place at a very rapid pace. During normal everyday conversation, native speakers go from building a clause structure to phonetic encoding in about 40 milliseconds (Van Turennout, Hagoort, and Brown, 1998). Similarly, the other stages shown in Figure 1 have to be performed within the small period of time available to a speaker involved in everyday communication. The typical window in turn taking is about 500-1000 milliseconds (Bull and Aylett, 1998). Although the majority of research on language production and comprehension has been conducted with English speakers, it is assumed that results closely resemble the processing of all languages including Spanish. If language users cannot perform the internal activities presented in Figure 1 in real-time, they will not be able to interact in communication as effective listener/speakers. Thus, spoken language facility is essential in successful oral communication.

Because the Versant Spanish Test is delivered in real time, it estimates the test taker's level of automaticity with the language. Automaticity 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, and Schriefers, 2003; Levelt, 2001). Automaticity is required for the speaker/listener to be able to devote attention to what needs to be said rather than to how the code is to be structured/analyzed. By measuring basic encoding and decoding of oral language as performed in integrating tasks in real time, the Versant Spanish Test infers the degree of automaticity in language performance.

Two basic types of scores are produced from the test: scores relating to the content of what a test taker says and scores relating to the manner of the test taker's speech. This distinction corresponds roughly to Carroll's (1961) description of a knowledge aspect of language performance and a control aspect. 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.

Some measures of automaticity can be misconstrued as memory tests. Since some of the tasks involve repeating long sentences or holding phrases in memory in order to piece them together into reasonable sentences, it may seem that these tasks measure memory instead of language ability. However, psycholinguists have shown that short-term or verbal working memory for such things as remembering a string of digits is distinct from cognitive resources used to process and comprehend sentences (Caplan & Waters, 1999). The fact that syntactic processing resources are generally separate from short-term memory stores is also evident in the empirical results of Versant Spanish Test validation experiments (see Section 5: Validation). Empirical findings show that virtually all native speakers achieve high scores on the test whereas non-native speakers obtain scores distributed across the scale. If memory, as such, were an important component of performance on the Versant Spanish Test, then the native speakers would show greater variation according to the range of memory spans. Also, the test would not correlate so highly with other accepted measures of oral proficiency as it does, since it would be measuring memory capacity and not language ability.

Note, however, that the Versant Spanish 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 test taker's facility with the language itself and is not confounded with other factors. The goal is to tease apart familiarity with spoken language from cultural knowledge, understanding of social relations and behavior, and the test taker'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 on collecting performance data that is directly relevant to the test construct.

The Versant Spanish Test measures the real-time encoding and decoding of spoken Spanish. Performance on Versant Spanish Test items predicts a more general spoken language facility, which is essential in successful oral communication. The same facility in spoken Spanish that enables a person to satisfactorily understand and respond to the listening/speaking tasks in the Versant Spanish Test also enables that person to participate in everyday native-paced Spanish conversation.

# 3. Content Design and Development

All Versant Spanish Test items were designed to be region neutral. The content specification also requires that both native speakers and proficient non-native speakers find the items very easy to understand and to respond to appropriately. For Spanish learners, the items cover a broad range of skill levels and skill profiles.

Each Versant Spanish Test item is independent of the other items and presents unpredictable spoken material in Spanish. 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 test taker's facility with the language itself. Thus, the test performance relates most closely to language abilities and is not confounded with other test-taker characteristics. Third, context-independent tasks maximize response density; that is, within the time allotted, the test taker has more time to demonstrate performance in speaking the language. Less time is spent developing a background cognitive schema needed for successful task performance.

## 3.1 Vocabulary Selection

The vocabulary used in the test items and responses was restricted to forms of the 8,000 most frequent words in the Spanish Call Home corpus of spontaneous spoken dialogues available from the Linguistic Data Consortium (LDC) at the University of Pennsylvania. The 8,000 most common roots were used to create the base lexicon for the Spanish test item development. A small number of other related words were also included for completeness; for example, if the word abierto (open) was included in the 8,000 roots in the lexicon, then the word cerrado (closed) might be allowed as well in order to create certain natural item texts.

## 3.2 Item Development

Versant Spanish Test items were drafted by an Argentine item developer. In general, the language structures used in the test reflect those that are common in everyday Spanish. The items were designed to be independent of social nuance and higher cognitive operations.

Draft items were then sent for outside review to ensure that they conformed to current colloquial Spanish usage in different countries. Dialectically distinct native Spanish speaking linguists reviewed the items to identify any geographic bias. The reviewers were from Chile, Colombia, Ecuador, Mexico, Puerto Rico, Spain, and Venezuela. All items, including anticipated responses for Short Answer Questions, were checked for compliance with the vocabulary specification. Vocabulary items that were not present in the lexicon were either changed to other entries that were listed or kept and added to a supplementary vocabulary list. The changes proposed by the different reviewers were then reconciled and the original items were edited accordingly.

Expert judgment was used initially to define correct answers to Short Answer Question items. Many of the items have multiple answers that are accepted as correct. All questions were pre-tested on diverse samples of native and non-native speakers. For an item to be retained in the test, it had to be understood and responded to appropriately by at least 80% of a reference sample of educated native speakers of Spanish.

## 3.3 Item Prompt Recordings

Native speakers from six different Spanish-speaking countries with clear voices were selected for recording all voice materials. The voice talents that recorded the test items were three females (from Peru, Puerto Rico, and Spain) and three males (from El Salvador, Mexico, and Spain). Also two examiner

voices were recorded: one from Latin America and the other from Spain. The examiner voices were distinct from the speakers used for the test items.

Recordings were made in a professional recording studio in Menlo Park, California. In addition to the Spanish instruction prompts, English prompts were recorded by an English examiner voice. Thus, there are two versions of the Versant Spanish Test – one with a Spanish examiner voice and the other with an English examiner voice.

The recordings were reviewed for quality and clarity. For final quality control of the audio files, the entire test was then given to ten native speakers of Spanish, including five Spanish interpreters. Any problematic recordings were removed.

# 4. Scoring

Of the 58 presented items in an administration of the Versant Spanish Test, 51 responses are used in the automatic scoring. The first item response in each part of the test (with the exception of Part F: story retellings), is considered a practice item and is not incorporated into the final score. In Part F, both story retellings items are scored. In addition, the two open questions in Part G are not scored automatically. The responses from Part G are available, however, for review by the test administrator.

The Versant Spanish Test score report is comprised of an Overall score and four diagnostic subscores (Sentence Mastery, Vocabulary, Fluency[1] and Pronunciation).

**Overall:** The Overall score of the test represents the ability to understand spoken Spanish and speak it intelligibly at a native-like conversational pace on everyday topics. Scores are based on a weighted combination of the four diagnostic subscores (30% Sentence Mastery, 20% Vocabulary, 30% Fluency and 20% Pronunciation). Scores are reported in the range from 20 to 80.

**Sentence Mastery:** Sentence Mastery reflects the ability to understand, recall, and produce Spanish phrases and clauses in complete sentences. Performance depends on accurate syntactic processing and appropriate usage of words, phrases, and clauses in meaningful sentence structures.

**Vocabulary:** Vocabulary reflects the ability to understand common everyday words spoken in sentence context and to produce such words as needed. Performance depends on familiarity with the form and meaning of everyday words and their use in connected speech.

---

[1] 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 Versant Spanish Test 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 Spanish Test 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.

**Fluency:** Fluency is measured from the rhythm, phrasing and timing evident in constructing, reading and repeating sentences.

**Pronunciation:** Pronunciation reflects the ability to produce consonants, vowels, and stress in a native-like manner in sentence context. Performance depends on knowledge of the phonological structure of everyday words as they occur in phrasal context.

Figure 2 illustrates which sections of the test contribute to each of the four subscores. Each vertical rectangle represents the response utterance from a test taker. The items that are not included in the automatic scoring are shown in blue. These include the first item in each of the scored sections of the test and the last task: Open Questions.
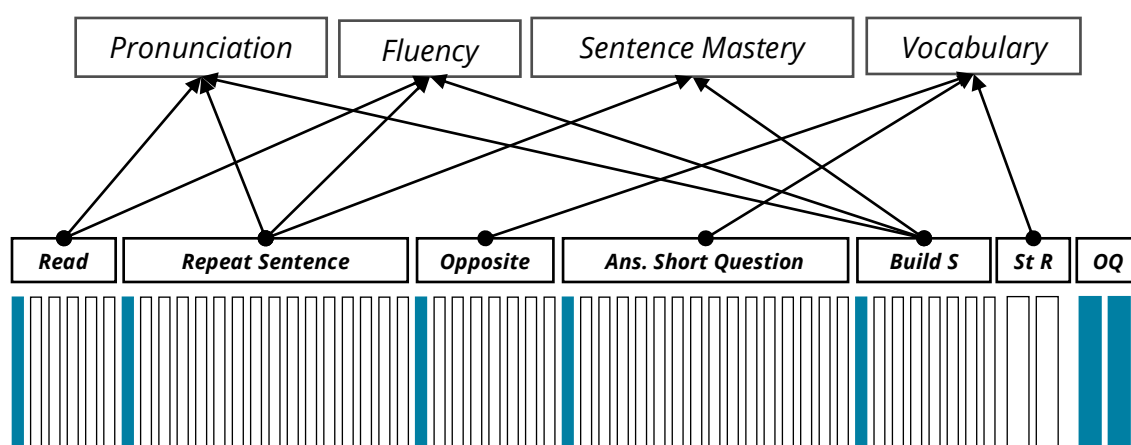


Figure 2. Relation of subscores to item types.

The subscores are based on two different aspects of language performance: a knowledge aspect (or the content of what is said), and a control aspect (or the manner in which a response is said). The four subscores reflect these aspects of communication where Sentence Mastery and Vocabulary are associated with content and Fluency and Pronunciation are associated with manner. The content accuracy dimension counts for 50% of the Overall score and indicates whether or not the test taker understood the prompt and responded with appropriate content. The manner-of-speaking scores count for the remaining 50% of the Overall score, and indicate whether or not the test taker speaks like a native (or like a favorably-judged non-native). 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. In the Versant Spanish Test scoring logic, content and manner (i.e., accuracy and control) are weighted equally because successful communication depends on both.

The machine automatically calculates the content subscores by using a speech recognizer that has been optimized for non-native speech. In this aspect, the machine generally recognizes words as well or better than a naïve listener, but not quite as well as a trained listener who knows the item content. The manner-of-speaking scores (fluency and pronunciation or 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, these measures were automatically generated on a sample set of utterances (from both native and non-native speakers) and then were recalibrated according to a model of human ratings. The human raters from the validation experiment (Section 5) consistently applied a set of criteria to rate test takers' pronunciation and fluency from the Reading, Repeats, and Sentence Builds speech files. These human scores were then used to rescale the machine scores so that the pronunciation and fluency subscores generated by the machine optimally predict the human ratings.

## 4.1 Score Use

Once a test taker has completed a test, the Versant testing system analyzes the spoken performances and posts the scores at www.pearson.com/versant. Test administrators and score users can then view and print out the test results from a password-protected section of the website.

Score users may be educational and government institutions as well as commercial and business organizations. Versant Spanish Test scores may be used for making valid decisions about oral Spanish interaction 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 Spanish Test themselves or by having trusted third parties administer the test. In several countries, education and commercial institutions provide such services.

Versant Spanish Test scores can be used to evaluate the level of spoken Spanish skills of individuals entering into, progressing through, and exiting Spanish language courses. Scores may also be used effectively in evaluating whether an individual's level of spoken Spanish is sufficient to perform certain tasks or functions requiring mastery of spoken Spanish.

The Versant Spanish Test score scale covers a wide range of abilities in spoken Spanish communication. Score users must decide what Versant Spanish Test score can be regarded as a minimum requirement in their context. Score users may wish to base their selection of an appropriate criterion score on their own localized research. Versant Benchmarking Kits help score users determine the appropriate criterion scores.

## 5. Validation

Data sets collected from over 1,000 test takers were analyzed in a series of validation experiments. Two kinds of analysis of the resulting data support the validity of the Versant Spanish Test as a measure of proficiency in spoken Spanish. First, a comparison of score distributions shows that native speakers are clustered at the high end of the score scale, whereas Spanish language learners are distributed across a wide range of scores. Second, machine-generated Versant Spanish Test scores are compared with scores from other well-accepted human-rated assessments of spoken Spanish and with scores assigned by expert raters after listening to recorded speech samples.

## 5.1 Method

### 5.1.1 Materials

Note that the Versant Spanish Test administered in the validation studies is the same as the Versant Spanish Test described above in this document, except that the tests administered in the validation studies included three Open Questions and three Story Retellings. In consequence of the extra time required to play two extra prompts and record two extra responses, the duration of the Versant Spanish Test in the validation studies was about 17 minutes, on average. Test takers did not take a practice test prior to the actual administration. (However, research conducted by the Versant Test Development team has demonstrated that performance on the Versant is not improved by taking a practice test.)

### 5.1.2 Participants

Three sets of participants were required for the validation experiments: a group of Spanish native speakers (as test takers), a group of non-native speakers (as test takers), and trained human raters to assess recorded speech samples and to conduct Oral Proficiency Interviews (OPIs).

**Native Speakers**
Adult native Spanish speakers (18 years old or older) were recruited for norm-referencing and test validation. Native speakers were roughly defined as individuals who spent the first twenty years of their lives in a Spanish-speaking country, were educated in Spanish through college level, and currently reside in a Spanish-speaking country. Samples were gender-balanced when possible. Four hundred twenty-two candidates constituted the native Spanish speaker sample: 135 from Argentina, 36 from Colombia, 217 from Mexico, 21 from Puerto Rico, and 13 from other Latin American countries. In addition, 153 native Spanish speakers from Spain were recruited for further validation, bringing the total to 575 native speakers.

**Non-Native Speakers**
For the non-native sample, the Versant Test Development team contacted a number of Spanish departments at universities in the United States asking them to have students take the Versant Spanish Test and, if possible, an official Spanish OPI certified by the American Council on the Teaching of Foreign Language (ACTFL). Students/universities were remunerated for their participation and for the ACTFL test fee. In addition, test takers were recruited from the military and other institutions. A subset of each group took one of two oral interview tests (ACTFL OPI or SPT-Interview-ILR; see below). A total of 574 non-native speakers participated in the experiment.

**Human Raters**
The validation experiments called for several groups of human raters to perform the proficiency interviews and to analyze spontaneous speech files collected from the last two tasks (Story Retellings and Open Questions) in the test. Two groups of raters conducted two types of oral interviews:

Raters from Language Testing International (LTI) administered all the certified OPIs for the ACTFL. The ACTFL OPI interviews were all conducted by telephone. LTI is the exclusive licensee for ACTFL testing. (The scores from these interviews will be referred to as ACTFL OPI.)

Government-certified raters administered a telephone OPI according to the Spoken Proficiency Test (SPT) procedure, and the scores were reported on the ILR scale. The two government-certified raters also worked under contract for the Federal Bureau of Intelligence (FBI) and had experience administering SPTs in Spanish for the government. Both of the government-certified raters were female; one was from Peru and the other was from Colombia. (Scores from these interviews will be identified as SPT-Interview-ILR.)

In addition, human raters were recruited to analyze a total of six 30-second recorded responses elicited by the Open Questions and Story Retellings at the end of each Versant Spanish Test. These six recordings are referred to as the 30-second response samples. Human ratings of the 30-second response samples were supplied by three rater sets:

Three native speakers of Spanish were selected to listen to the 30-second response samples and assign level descriptors to them based on the CEFR. All had degrees from universities in South America. Two of the raters were certified Spanish translators/interpreters. Raters received training in the CEFR level rubrics prior to engaging in the rating tasks and were tracked during the rating process to ensure that they were following defined rubrics. Training included rating subjects not used in the main study until a predetermined level of agreement was reached. (The scores generated from these raters will be referred to as CEFR Estimates.)

Four government-certified ILR interviewers listened independently to the six 30-second response samples from each of 166 test takers. Two raters were active in testing at the Defense Language Institute (DLI), Monterey, California, and provided estimated ratings based on the ILR scale descriptors. (These ratings will be identified as ILR-Estimate/DLI.)

The other two government-trained raters were the same two people who administered the SPT-Interview-ILR. After a pause of two weeks, these raters also listened independently to the 30-second response samples from the Versant Spanish Test administrations and provided estimated ratings based on the ILR scale. (These ratings will be referred to as ILR-Estimate/SPT.) These 996 (6 response samples from 166 test takers) 30-second response samples included the samples from the 37 test takers that they had interviewed.

### 5.1.3 Experimental Design

For the validation experiments, both the native Spanish speakers and the non-native speakers took the Versant Spanish Test. All native speakers took the test in their home countries. The collected data consisted of automatically generated Overall scores and subscores and spontaneous 30-second response samples from the Open Questions and Story Retellings.

The Versant Spanish Test Overall scores for native Spanish speakers were compared with the scores of non-native speakers and were correlated with various human ratings. Several sets of human scores were collected for this purpose:

### Ratings on oral language assessment tests

**ACTFL OPI** - For the ACTFL interviews, the Versant Test Development team coordinated with universities to have their students take an ACTFL interview within a day of the Versant Spanish Test. The standard ACTFL interview was administered with at least two official ACTFL ratings per interview. ACTFL submitted 52 scores, one for each of the 52 participants.

**SPT-Interview-ILR** - For the SPT interviews, the candidates were asked to take the Versant Spanish Test within one day of the SPT interview. The raters followed the ILR level descriptions that appear in the Test Manual of the Speaking Proficiency Test developed by the Federal Language Testing Board. Each rater independently provided ILR-based proficiency level ratings for each of the 37 candidates, for a total of 74 ratings.

### Ratings on 30-second response samples

**CEFR-Estimate ratings** – CEFR ratings were based on the CEFR level descriptors. For the CEFR Estimates, 30-second response samples from 572 test takers were rated. On average, the three raters together provided 11 independent scores for each test taker, resulting in a total of 6,125 ratings.

**ILR-Estimate/DLI ratings** – These ratings were based on the ILR/OPI rubrics. Nine hundred ninety-six 30-second response samples from 166 test takers were scored. The two raters provided a total of 1978 ratings, with an average of 12 independent scores for each test taker.

**ILR-Estimate/SPT ratings** - These ratings were based on the rubrics in the Tester manual of the Speaking Proficiency Test developed by the Federal Language Testing Board. Nine hundred ninety-six 30-second response samples from 166 test takers were scored. On average, the two raters provided 19 independent scores for each test taker, resulting in 2,798 ratings total.

For all the CEFR and ILR estimates, raters based their independent judgments on the recorded spoken responses to the Versant Spanish Test Open Questions and Story Retellings. Six 30-second response samples were rated for each test taker. Figure 3 shows the detailed design including the number of test takers who contributed scores for the Versant Spanish Test and for each interview.
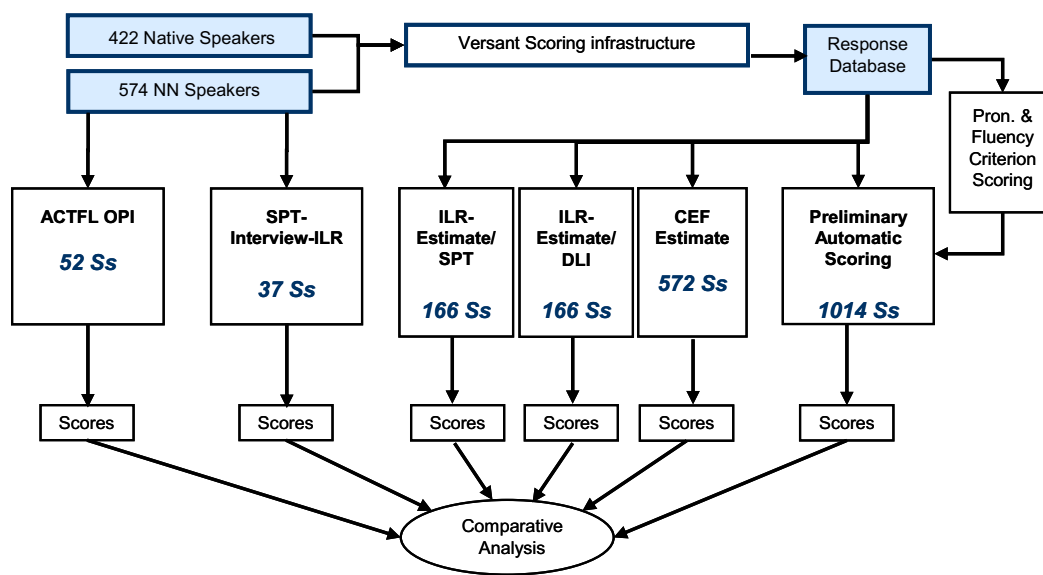
Figure 3. Design for validity experiment.

## 5.2 Results

There were two goals of the validation experiments. One preliminary goal was to determine if native speakers obtain high scores on the Versant Spanish Test, while speakers of other languages who are learning Spanish are distributed over a wide range of scores. If this expected distinction between the two known populations holds, it may be seen to support the Versant Spanish Test's validity. The principal goal of the experiments was to understand the relation of Versant Spanish Test scores to the scores obtained using well-documented human-mediated measures of oral proficiency and expert human estimates of test-taker performance using the well-known ILR proficiency scale.

### 5.2.1 Test Reliability

The split-half reliability of the Overall score of the Versant Spanish Test is 0.96 (N = 267, with a standard error of measurement of 2.6). The reliability represents the corrected calculation due to split-half underestimation. The high reliability score is a good indication that the computerized assessment will be consistent for the same test taker assuming no changes in language proficiency. Table 2 lists the split-half reliabilities for each of the subscores.

Table 2. Split-half Reliability for Versant Spanish Test Scores (N = 267)

| Score | Reliability |
|---|---|
| Overall | .97 |
| Vocabulary | .91 |
| Pronunciation | .94 |
| Sentence Mastery | .95 |
| Fluency | .93 |

## 5.2.2 Native and Non-native Group Performance

Figure 4 presents cumulative distribution functions for two speaker groups: natives and non-native speakers of Spanish. The distribution of the native speakers clearly distinguishes the natives from the non-native sample. For example, fewer than 5% of the native speakers score below 75, while only 10% of the non-native speakers score above 75. The Overall score and the subscores are reported in the range from 20 to 80 (scores above 80 are reported as 80). Thus, in Figure 4, the trajectory of the curve from the Overall score of 80 and above is presented in a gray area.
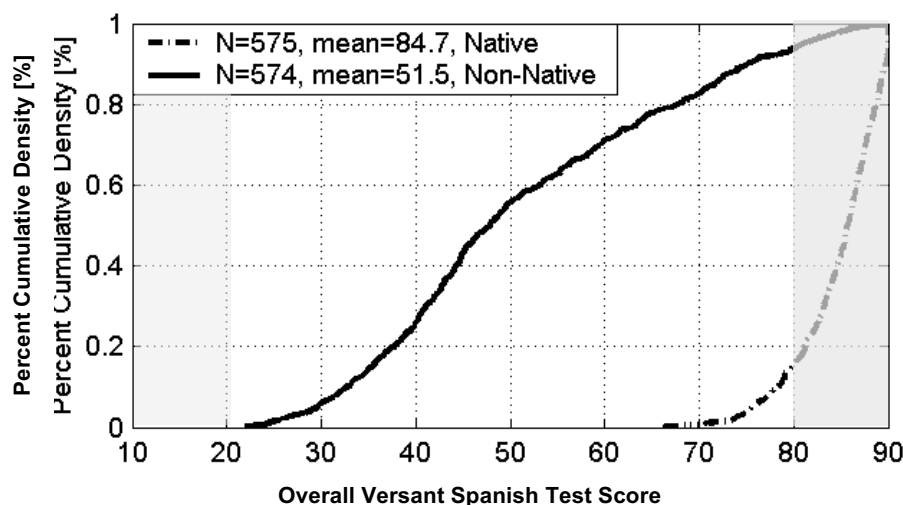


Figure 4. Cumulative distribution functions for native and non-native speakers.

This finding suggests that the Versant Spanish Test has high discriminatory power among learners of Spanish as a second or foreign language, whereas native speakers obtain near-maximum scores. Further analysis has shown that the same patterns hold true regardless of gender or age.

Since the test was normalized with Spanish speakers from Latin American countries, a further validation experiment was conducted to investigate whether or not Iberian Spanish speakers perform comparably to other natives. The cumulative distribution functions of native Spanish speaking test takers from five different countries are presented in Figure 5.
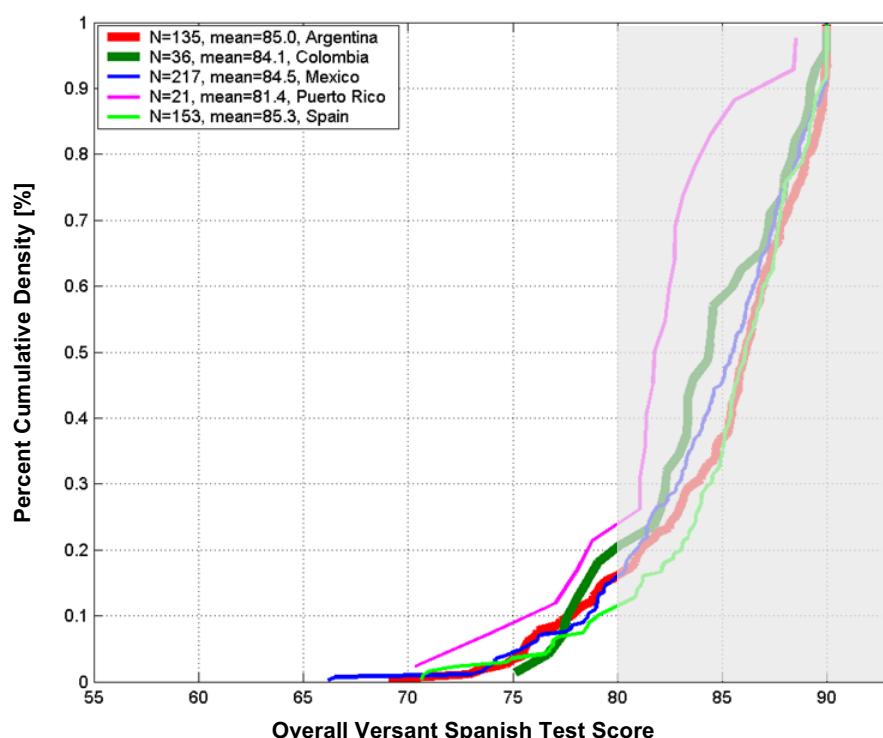
Figure 5. Cumulative distribution functions for native Spanish speakers of different countries.

As seen in Figure 5, the Iberian Spanish speakers perform equally as well as speakers from the Latin American countries. The findings support the notion that the Versant Spanish Test scores reflect a speaker's facility in spoken Spanish irrespective of the speaker's country of origin.

### 5.2.3 Correlations between the Versant Spanish Test and Human Scores

A main goal of the validity experiments was to understand how the scores automatically generated for the Versant Spanish Test correlate with other measures of spoken Spanish from trained human raters. Certified OPIs (human-administered and human-rated) are well accepted as valid measures of spoken communication skill, or oral proficiency. The first step in the analyses was to confirm that the human raters were reliable in their scoring. This is an important step in showing that the scores used in the validation experiments were consistent.

For the SPT-Interview-ILR, the inter-rater reliability was r=0.93. No direct reliability information was reported for the ACTFL OPI. Reliability of the CEFR Estimate ratings was $r = 0.92$. For the estimated ILR-Estimate/DLI, the reliability was 0.95, and for the ILR-Estimate/SPT, it was $r = 0.97$.

Together, the reliability data of the human raters indicate that the selected human raters are internally reliable.

Note that inter-rater reliability is not applicable to machine methods. Any machine running the same software will produce exactly the same results for the same signal each time it is analyzed, leading to an inter-rater reliability of 1.0.

## 5.2.4 Validity of Scores

Results from the Versant Spanish Test were correlated with two different human-conducted and human-rated OPIs. As mentioned above, these were the ACTFL OPI and the SPT-Interview-ILR.

For one part of the data analysis, a one-parameter Rasch model as implemented in the computer program FACETS (Linacre, 2003) was used. The FACETS program estimates rater severity, subject ability, and item difficulty (Linacre, Wright, and Lunz, 1990). The model assumes a single underlying dimension (i.e., proficiency in spoken Spanish), where values along this dimension are expressed in mathematical units called Logits. For all the human ratings displayed in Figures 6-10 below, the boundaries of the different levels were mapped onto a continuous Logit scale, and displayed as such.

Figure 6 presents a scatter plot of the ACTFL OPI ratings as a function of Versant Spanish Test scores for 52 Spanish learners.
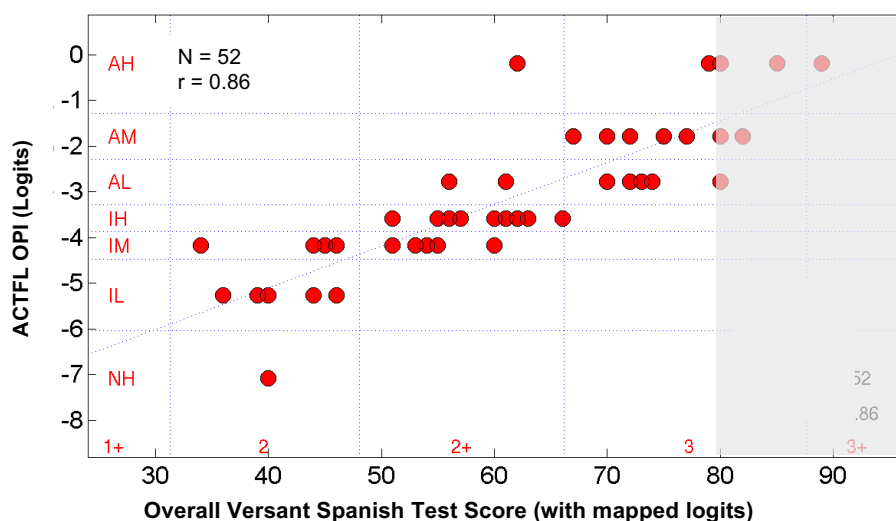


Figure 6. ACTFL OPI ratings as a function of Versant Spanish Test scores. The Y-axis displays the ACTFL OPI levels in logits as well as their corresponding titles: Novice High (NH), Intermediate Low (IL), Intermediate Middle (IM), Intermediate High (IH), Advanced Low (AL), Advanced Middle (AM), and Advanced High (AH).

The correlation for these two tests is 0.86, indicating a strong relation between the machine-generated scores and the human-rated interviews.

The other comparison of the machine scores with an oral proficiency interview was between the Versant Spanish Test and the SPT-Interview-ILR. Figure 7 shows the scatter plot of ratings for 37 non-native Spanish speakers for these two measures.
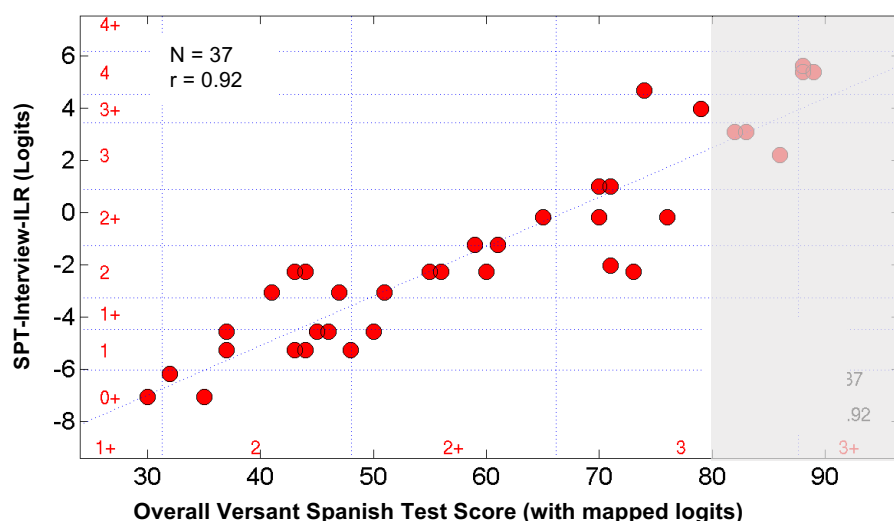
Figure 7. SPT-Interview-ILR scores as a function of Versant Spanish Test ratings.

The correlation between the Versant Spanish Test scores and the ratings from the SPT-Interview-ILR is 0.92. Like the ACTFL versus the Versant Spanish Test data set, the SPT versus the Versant Spanish Test data set contains no outliers. Together, the close alignment between the Versant Spanish Test Overall scores and the scores from these other two certified tests of oral proficiency indicates that the Versant Spanish Test produces scores that correspond with other accepted methods of assessment.

Additional validation analyses involved comparing the Versant Spanish Test scores with human ratings of the 30-second response samples collected during the administration of the Versant Spanish Test (responses to the Open Questions and the Story Retellings). The independent estimates of the test takers' spoken Spanish skill according to the CEFR and ILR levels can provide additional evidence of validity for the Versant Spanish Test. The speech samples were analyzed by three different pairs of human raters, one pair using the CEFR scale, and two pairs using the ILR scale.

The first comparison was the Versant Spanish Test scores with CEFR Estimates. Figure 8 is a scatter plot of results on these two measures for 572 participants.
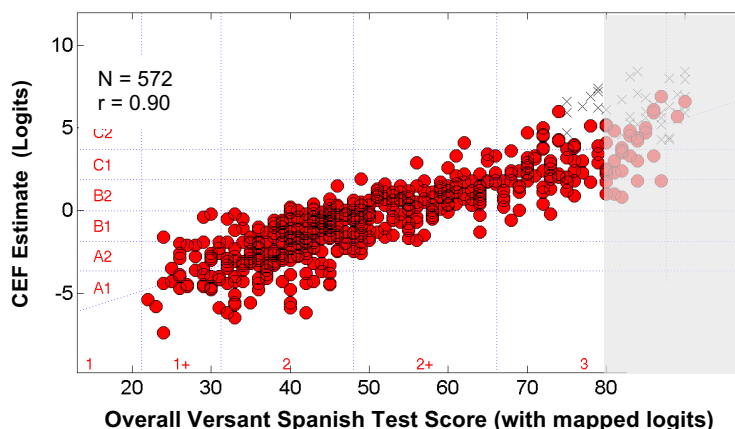


Figure 8. CEFR Estimates on spontaneous speech samples as a function of Versant Spanish Test scores.

The Versant Spanish Test scores are highly correlated with the CEFR Estimates ($r = 0.90$). This data set also contains no obvious outliers, indicating a close match between the machine-generated Versant Spanish Test scores and the scores from the human raters. Note that this data set included 42 native speakers of Spanish. In Figure 8, the native speaker data points are represented by Xs, while the Spanish learners are filled circles, as in the other Figures 6-10.

The next comparison was the Versant Spanish Test scores with ILR-Estimate/DLI ratings for spontaneous speech samples collected from the Versant Spanish Test for 166 participants. Figure 9 illustrates the strong correlation ($r = 0.88$) between the two measures.
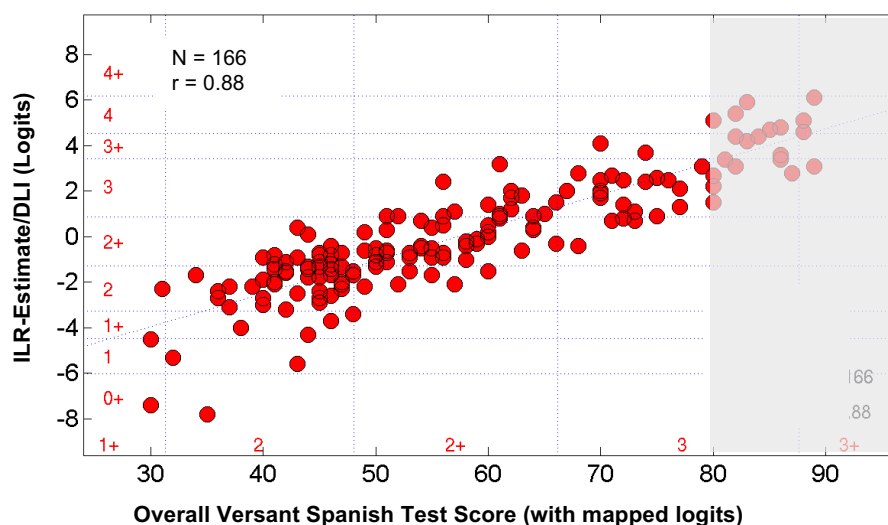
Figure 9. ILR-Estimate/DLI on spontaneous speech samples as a function of Versant Spanish Test scores.

The final comparison was between the Versant Spanish Test scores and ILR-Estimate/SPT ratings. Figure 10 shows the scatter plot for 166 non-native test takers.
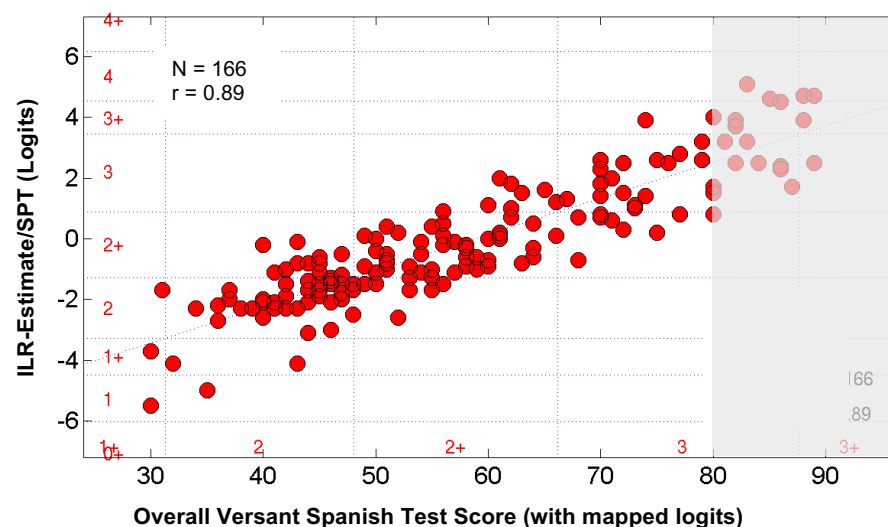
Figure 10. ILR-Estimate/SPT on spontaneous speech samples as a function of Versant Spanish Test scores.

The correlation between the Versant Spanish Test scores and ILR-Estimate/SPT ratings is 0.89, with no outliers evident.

From these results, it is clear that all three comparisons of the Versant Spanish Test scores with estimates based on other rubrics of oral proficiency show a close relation between the machine-generated scores and the human ratings of speech samples from the same test takers.

Table 3 summarizes all the correlation results above in addition to other comparisons between the human rated scores. Some of the comparisons are not possible because the data do not overlap, for example, no respondent participated in two oral proficiency interviews, such as the ACTFL OPI and the SPT-Interview-ILR.

Table 3. Correlations between different measures of oral proficiency.

| | Versant Spanish Test | ACTFL OPI | SPT-Interview-ILR | CEFR Estimate | ILR-Estimate/DLI | ILR-Estimate/SPT |
|---|---|---|---|---|---|---|
| Versant Spanish Test | | 0.86 | 0.92 | 0.90 | 0.88 | 0.89 |
| ACTFL OPI | 0.86 | | -- | -- | 0.87 | 0.89 |
| SPT-Interview-ILR | 0.92 | -- | | -- | 0.92 | 0.94 |
| CEFR Estimate | 0.90 | -- | -- | | 0.91 | 0.91 |
| ILR-Estimate/DLI | 0.88 | 0.87 | 0.92 | 0.91 | | 0.96 |
| ILR-Estimate/SPT | 0.89 | 0.89 | 0.94 | 0.91 | 0.96 | |

The Versant Spanish Test and other assessments of oral proficiency are significantly related with correlation coefficients ranging from 0.86 to 0.92 (all $p<0.01$). In addition, all the correlations between the oral proficiency interviews and the Versant Spanish Test are nearly identical to the interview's correlation with human-rated estimates of other measures. For example, in Table 3, the Versant Spanish Test scores correlate with the ILR scores from the SPT-Interview-ILR with a correlation coefficient of 0.92, while the ILR estimates from the two pairs of certified ILR interviewers correlated with the actual SPT-Interview-ILR scores with coefficients of 0.92 and 0.94. Thus, the Versant Spanish Test procedure elicits sufficient spoken language behavior on which to base a reliable and accurate human judgment of practical speaking and listening skills. The Versant Spanish Test also produces results that are similar to those of expert human raters.

## 5.3 Item Pool Expansion

More recently, the item pool of the Versant Spanish Test was increased by 150%. Items were drafted and validated in the same manner as described above. Seven hundred thirty-nine native speakers from Spain, Costa Rica, Mexico, Ecuador, Columbia, Peru, Argentina, and Puerto Rico served as models for these responses. Eight hundred twenty-three non-natives, representing a wide range of native

languages and spoken Spanish abilities, also provided responses to these new items. The speech-recognition models were then re-trained on this data.

To assure that the re-trained models were comparable to the previous models, a randomly selected sub-section (N = 268) of the above described participants (10% natives, 90% non-natives) were used to compare scoring from the previous model to the current model. The data from these participants was not used in re-training the models. Comparison of the scores generated from the previous model and those from the re-trained models demonstrated a nearly perfect correlation (*r* = .995) with the previous model (see Figure 11). This result suggests that the data, inferences, and validity evidence based on the previous models remain warranted.
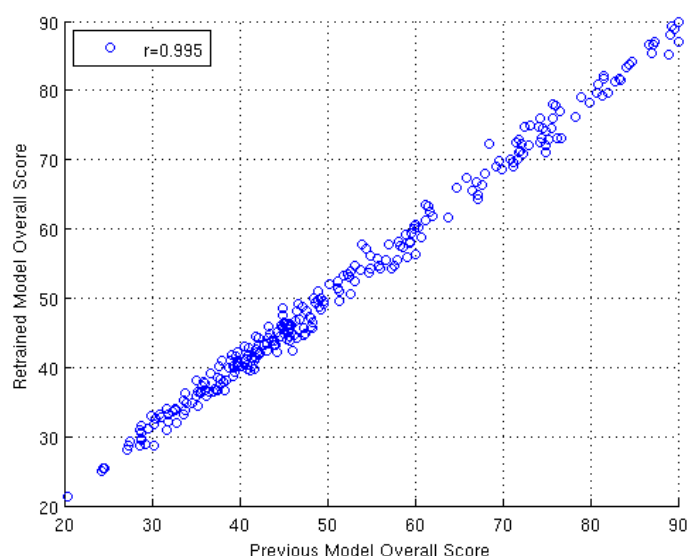


Figure 11. Correlation of overall Versant Spanish Test using scoring from previous model vs. re-trained model.

## 5.4 Conclusions

The validation data indicate that native speakers consistently obtain high scores on the Versant Spanish Test, while learners of Spanish are distributed over a wide score range. This separation of the two sample populations illustrates the power of the Versant Spanish Test as a measurement instrument of spoken language ability.

In addition, the validation experiments show that Versant Spanish Test scores are reliable and correlate strongly with other measures of spoken Spanish. Generally, the Versant Spanish Test scores will accurately predict 80% or more of the variance in US government oral language proficiency ratings. Given the fact that the reliability of ILR scales is reported in the literature to be around 0.90 (Stansfield & Kenyon, 1992), this is as high as the predictive power between two consecutive independent oral proficiency interviews. The Versant Spanish Test also predicts ACTFL interview scores with the same reliability as two independent ACTFL OPIs predict one another (Sieloff-Magnan, 1987).

The validation results show that the Versant Spanish Test is both a reliable and valid measure of test takers' facility with spoken Spanish and can be used confidently in assessing the spoken language abilities of non-native speakers of Spanish.

To assure the defensibility of employee selection procedures, employers in the U.S. follow the Equal Employment Opportunity Commission's (EEOC's) Uniform Guidelines for Employee Selection Procedures. These guidelines state that employee selection procedures must be reliable and valid. The above information provides evidence of the reliability, validity and legal defensibility of the Versant Spanish Test in conformance with the prescriptions of the EEOC's Uniform Guidelines.

# 6. About the Company

**Ordinate Testing Technology:** The patented Versant automated testing system was developed to apply advanced speech recognition techniques and data collection via the telephone to the evaluation of language skills. The system includes automatic telephone reply procedures, dedicated speech recognizers, speech analyzers, databanks for digital storage of speech samples, and scoring report generators linked to the Internet. The Versant Spanish Test 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 human rating of spoken language samples.

**Pearson:** Ordinate Corporation, creator of the Versant tests, was combined with Pearson's Knowledge Technologies group in January, 2008. The Versant line of tests are the first to leverage a completely automated method for testing spoken language.

**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 at investigating new applications for Ordinate technology. Research results are published in international journals and made available through the Versant test website (www.pearson.com/versant).

# 7. References

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 Converence on Spoken Language Processing. Canberra, Australia: Australian Speech Science and Technology Association.

Caplan, D. & Waters, G. (1999). Verbal working memory and sentence comprehension. Behavioral and Brain Sciences, 22, 77-126.

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.

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.

Linacre, J.M. (2003). Facets Rasch measurement computer program. Chicago: Winsteps.com.

Linacre J.M., Wright B.D., & Lunz, M.E. (1990). A Facets model for judgmental scoring. Memo 61. MESA Psychometric Laboratory. University of Chicago. www.rasch.org/memo61.htm.

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.

Sieloff-Magnan, S. (1987). Rater reliability of the ACTFL Oral Proficiency Interview, Canadian Modern Language Review, 43, 525-537.

Stansfield, C.W., & Kenyon, D.M. (1992). The development and validation of a simulated oral proficiency interview. Modern Language Journal, 76, 129-141.

Van Turennout, M. Hagoort, P., & Brown, C.M. (1998). Brain activity during speaking: From syntax to phonology in 40 milliseconds. Science, 280, 572-574.

# 8. Appendix: Test Paper

**Side 1 of the Test Paper:** Instructions and general introduction to test procedures. **Note:** These instructions are available in several different languages.

**Side 2 of the Test Paper:** Individualized test form (unique for each test taker) showing Test Identification Number, Part A; sentences to read, and examples for all sections.

## 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.pearson.com/versant

Version 1118M