



Versant™ Junior English Test – Level 2

Test Description and Validation Summary

Table of Contents

1. Introduction	3
2. Test Description	3
2.1 Test Design	3
2.2 Test Administration	3
2.2.1 Telephone Administration	4
2.2.2 Computer Administration	4
2.3 Test Format	4
Part A: Read the Other Word	4
Part B: Reading	5
Part C: Repeat	5
Part D: Questions	5
2.4 Number of Items	6
2.5 Test Construct	6
3. Content Development	7
3.1 Vocabulary Selection	7
3.2 Item Development	8
3.3 Field Testing	8
4. Scoring	9
4.1 Scoring and Weights	9
4.2 Score Use	11
5. Validation	12
5.1 Validation Study Design	12
5.2 Reliability	12
5.3 Validity	13
5.3.1 Content Validity	13
5.3.2 Construct Validity	14
5.3.3 Accuracy of Machine Scores	15
6. Conclusions	16
7. About the Company	17
8. References	17
9. Appendix: Test Materials	19

1. Introduction

The Versant Junior English Test (VJET) – Level II, powered by Ordinate technology, is an assessment instrument designed to measure how well a young English language learner understands and speaks basic English. The VJET Level II offers two levels to choose from depending on the candidate’s stage of learning: VJET Level I and VJET Level II. Both VJET tests are intended for children and take approximately 15 minutes to complete. Because the VJET tests are delivered automatically by the Ordinate testing system, the tests can be taken at any time, from any location by phone or via computer and a human examiner is not required. The computerized scoring allows for immediate, objective, and reliable results.

Both VJET Level I and Level II are designed to measure *facility* in basic spoken English. Facility is how well the young learner can understand spoken English on everyday topics and respond appropriately at a native-like conversational pace in English. As measured in the VJET tests, facility with spoken English is a key element in determining whether young English learners can be functional in spoken English performance.

2. Test Description

2.1 Test Design

During test administration, the Ordinate testing system presents a series of spoken prompts which have been recorded in English at a conversational pace and elicits oral responses in English. The voices that present the item prompts belong to native speakers of English, providing a range of speaking styles.

The VJET Level II has four sections: Read the Other Word, Reading, Repeat, and Questions. Each section elicits responses from the test-taker that are analyzed automatically by the Ordinate scoring system. The item types provide fully independent measures that underlie facility with spoken English, including phonological fluency, sentence construction and comprehension, passive and active vocabulary use, listening skill, and pronunciation of rhythmic and segmental units. Because more than one task type contributes to each subscore, the use of multiple item types strengthens score reliability.

The VJET Level II score report is comprised of an Overall score and five diagnostic subscores:

- Sentence Structure
- Word Recognition
- Word Meaning
- Fluency
- Pronunciation

Together these scores describe the test-taker’s facility in spoken English. The Overall score is a weighted average of the five subscores on a scale from 75 to 150.

The Ordinate testing system automatically analyzes the test-taker’s responses and posts scores on its website within minutes of completing the test. Test administrators and score users can view and print out test results from a password-protected section of Pearson’s website: www.VersantTest.com

2.2 Test Administration

Administration of the VJET Level II generally takes about 15 minutes over the phone or via computer. Regardless of the mode of test administration, it is best practice for the administrator to give the test

instructions and test paper to the test-taker at least five minutes before starting the test. The test-taker then has the opportunity to read the test instructions and test paper and ask questions before the test begins. The administrator should answer any procedural or content questions that the test-taker may have prior to beginning the test.

2.2.1 Telephone Administration

For telephone delivery, a corded, land-line telephone is required for accurate test results. Telephone administration is supported by a test paper. The test paper presents the telephone number to call and a Test Identification Number (TIN) that is unique for each test administration. When the telephone number is dialed, the Ordinate testing system prompts the user to enter the TIN on the telephone keypad. During the test, the instructions and examples are spoken by an examiner voice and are written verbatim on the test paper. The test instructions and examples are available in English, Chinese, Korean, Japanese, and Spanish. The test items are spoken in English by a range of native English speakers distinct from the examiner voice. Test-takers interact with the test system in English until they complete the test and hang up the telephone.

2.2.2 Computer Administration

For computer delivery, the test-taker is fitted with a microphone headset connected to a computer. The computer must have an internet connection and Ordinate's Computer-Delivered Test (CDT) software (available at <http://www.versanttest.com/technology/platforms/cdt/index.jsp>). As in telephone delivery, the test-taker is provided with a unique TIN. Before the test begins, the system guides the test-taker through adjusting the volume and calibrating the microphone. 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 English, speaking their responses into the microphone headset. When the test is finished, the test-taker clicks a button labeled, "End Test".

2.3 Test Format

The following subsections provide brief descriptions of the task types and the abilities required to respond to the items in each of the four parts of the VJET Level II. The mechanism for the delivery of the recorded item prompts is interactive – the system detects when the test-taker has finished responding to one item and then presents the next item.

Part A: Read the Other Word

In this task, the test-taker sees a line of three words on the test paper or on the computer screen. The test-taker hears two of the three words and is asked to read aloud the printed word that he or she did not hear. (Note: The words in quotation marks are printed here for demonstration purposes only and are not printed on the actual test paper or displayed on the computer screen).

Examples:

1.	rose	dog	teacher	"dog teacher"
2.	trunk	pocket	like	"trunk like"
3.	pickle	leaves	apple	"pickle leaves"

By reading aloud the other word, the test-taker demonstrates recognition of spoken words and basic word-level reading skills. This task probes the test-taker's knowledge of letter-sound correspondence and word recognition independently from word meaning.

Part B: Reading

In this task, the test-taker reads aloud printed, numbered sentences, one at a time.

Examples:

1. That's my dog.
2. We walked in the park.
3. He likes to go to the movies every weekend.

The sentences are relatively simple in structure and vocabulary, and they can be read easily and in a fluent manner by literate, native English speakers. This task provides samples of the test-taker's pronunciation and reading fluency.

Part C: Repeat

In this task, the test-taker listens to a sentence, which is not printed on the test paper or displayed on the computer screen, and then is asked to repeat the sentence aloud verbatim. The sentences are presented in order of increasing difficulty.

Examples:

"Watch your step."
"I don't know what's in the box."
"If you keep going, the road will end at a big oak tree."

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). If a person habitually processes three-word phrases as a unit (e.g. "the red apple"), then that person can usually repeat utterances that contain several units of approximately the same 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 those who are not familiar with English words, phrase structures, and other common, meaningful language forms. Because the items in the Repeat section require test-takers to organize speech into linguistic units, it tests their familiarity with and mastery of English sentence structure. In addition, because test-takers are asked to repeat full sentences (as opposed to just words and phrases); the task also provides a sample of the test-taker's pronunciation and fluency in continuous spoken English.

Part D: Questions

In this task, the test-taker listens to a spoken question while looking at three printed words. The test-taker chooses the word that best answers the question and says the word aloud.

The questions are intended to be within the realm of familiarity of a native English-speaking 11-year-old, as determined by expert judgment and item statistics. The items do not presume any knowledge of specific facts of English or American culture, geography, history, or other subject matter. (Note: The questions are printed here for demonstration purposes only and are not printed on the actual test paper or displayed on the computer screen).

Examples:

1.	cat	book	chair	“Which one is an animal?”
2.	picture	gym	tree	“Which one has leaves?”
3.	teacher	clown	jogger	“Who has a painted face?”

To respond to the questions, the test-taker must be able to segment the words from a stream of continuous speech, extract meaning from the words in the question, and then must decode the printed words, access the word meanings, read individual words, and identify which best answers the question. The questions manifest a test of receptive vocabulary within the context of spoken questions presented in a conversational style.

2.4 Number of Items

The VJET Level II presents a total of 48 items to each test-taker over the four separate sections. Table I shows the number of items presented in each section.

Table I. Number of items presented per task.

Task	Presented
A. Read the Other Word	8
B. Reading	10
C. Repeats	14
D. Questions	16
Total	48

All items are drawn semi-randomly from a larger item pool. For example, each test-taker is presented with 8 *Read the Other Word* items that are selected randomly from those items available in the pool. This way, most items will be different from one test administration to the next. Proprietary algorithms are used by the Ordinate testing system to select from the item pool. These algorithms take into consideration, among other things, the item’s level of difficulty and the order of presentation. There are also four different fixed form tests available in which the items will not change from one test administration to the next. These fixed form tests are appropriate for purposes such as practice, placement, or progress monitoring.

2.5 Test Construct

The construct of a test is the concept or characteristics that a test is designed to measure (*Standards for Educational and Psychological Testing*, 1999). The VJET Level II is designed to measure *facility in spoken English* – that is, the ability to understand spoken English on everyday topics and to respond appropriately at a fully-functional conversational pace in intelligible English.

In order for a person to successfully participate in a spoken conversational exchange, numerous cognitive tasks must be completed very rapidly: a person must 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).

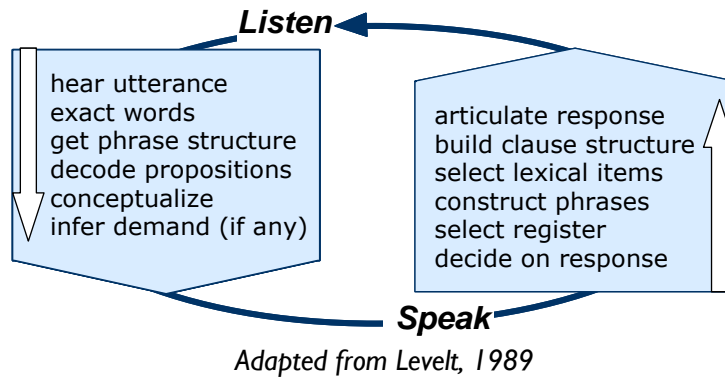


Figure 1. Conversational processing components in listening and speaking.

Core language component processes, such as lexical access and syntactic encoding, normally occur at a very rapid pace. All of the 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). If English language learners cannot perform the internal activities presented in Figure 1 in a rapid time frame, then they will not be able to interact in communication as effective listeners or speakers. Thus, spoken language facility is essential in successful oral communication.

Because the VJET Level II is designed to elicit responses from the test-taker 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). An example of automaticity is when an experienced driver is driving a car: the driver performs numerous operations simultaneously without having to think about them too much. Automaticity is required for a speaker/listener to be able to devote attention to what needs to be said rather than to how the message is to be structured and analyzed. By measuring basic encoding and decoding of oral and written language in real time, the VJET Level II probes the degree of automaticity in English language performance. The same facility in spoken English that enables an English language learner to satisfactorily understand and respond to the VJET Level II listening, speaking, and reading tasks also enables that learner to participate in everyday native-paced English conversation.

3. Content Development

3.1 Vocabulary Selection

The vocabulary used in the test items and responses was based on approximately 8000 of the most frequent lemmas (word forms) of a large sample of spontaneous spoken English available from the Linguistic Data Consortium (LDC) at the University of Pennsylvania.

Each VJET Level II item is independent of the other items and presents unpredictable spoken and written material in English. For English language learners, the items cover a broad range of skill levels and skill profiles. In general, the language structures used in the test reflect those that are common in everyday English. The items were designed to be independent of social nuance and high-cognitive functions.

3.2 Item Development

Items were drafted by a group of language teachers and curriculum designers based in Asia. Items were reviewed by a different group of English language experts to ensure that the items conformed to current colloquial English usage. Dialectically distinct native English speaking linguists reviewed the items to identify any geographic bias. All items, including anticipated responses, were checked for compliance with the vocabulary specification and non-compliant items were either edited or excluded altogether.

Native English speakers from the United States were selected to record all items. The item voices were both male and female, representing a range of ages and regional accents. The spoken test instructions and examples were recorded by different speakers and are available in English, Chinese, Korean, Japanese, and Spanish and are presented verbatim on the test paper or computer screen. Different examiner voices recorded the prompts for VJET Level I.

3.3 Field Testing

Each item was field tested with both native and non-native speakers of English. To ensure construct validity, the item specification requires that native speakers find the items very easy to understand and to respond to appropriately; this ensures that the item is measuring language ability rather than another construct. Native speakers were roughly defined as individuals who were born in an English-speaking country, were being educated in English, and currently resided in an English-speaking country. For the sample of non-native English speakers, participants representing a range of native languages and English proficiency levels were recruited. Young test-takers were sampled to ensure that the items would be appropriate for children. Table 2 presents the demographics of the native and non-native samples.

For an item to be included in the test, 90% of the native speaker sample had to provide a correct response to the item; otherwise, the item was eliminated.

Table 2. Demographics of the native and non-native samples.

Group	N	Male	Female	Gender Unknown	Age Range	Average Age	Native Languages
Native	345	169 (49%)	176 (51%)	0	8 to 17	14	English
Non-native	709	356 (50%)	350 (49%)	3 (1%)	9 to 18	14	186 Taiwanese (26%) 149 Korean (21%) 128 Japanese (18%) 122 Dutch (17%) 121 Spanish (17%) 3 Other (1%)

Item difficulty estimates were determined using a one-parameter Rasch model as implemented in the computer program FACETS (Linacre, 2003). The FACETS program estimates rater severity, ability level of the test-takers, and item difficulty (Linacre, Wright, and Lunz, 1990). The point biserial correlation between item response and total score was used to determine whether or not an item differentiated abilities well. Only items with a point biserial above a predetermined threshold were classified as gradable and became a part of the final test.

The data were also used to create language models and pronunciation and fluency models for automatic scoring. The acoustic models for the speech recognizer were optimized to accommodate for characteristics specific to children’s voices and reading and speaking patterns.

4. Scoring

4.1 Scoring and Weights

Scores for each test level consist of an Overall score and five diagnostic subscores: Sentence Structure, Word Recognition, Word Meaning, Fluency, and Pronunciation.

Overall: The Overall score is a measure of basic spoken English skills including recognition of spoken words and their meanings, mastery of English usage and sentence structure, and the ability to pronounce English in fluent sentences.

Sentence Structure: Sentence Structure is the level of mastery of grammar and usage in sentences.

Word Recognition: Word Recognition is the ability to identify words in spoken and written form.

Word Meaning: Word Meaning is the ability to understand common words in sentence context.

Fluency: Fluency is rhythm and timing in reading aloud and in repeating sentences.

Pronunciation: Pronunciation is the ability to say words intelligibly in sentence context.

Scores are reported on a scale from 75 to 150. The score report consists of an overall score and subscores, current capabilities, and to-improve statements.

The Overall score is calculated as a weighted combination of the five subscores. The Ordinate patented automated scoring system produces multiple, independent measures from the same set of responses. For this reason, tasks often contribute to more than one subscore. The use of multiple item types for subscores ensures score reliability. An added advantage of evaluating language skills independently is that a subscore is not confounded by features of other language skills. For example, a lack of fluency will not affect the evaluation of the content of the test-taker’s response. Figure 2 illustrates which sections of the test contribute to which subscores. Each vertical rectangle represents a single response utterance from a test-taker. The shaded vertical rectangles represent items that are not included in the automatic scoring (i.e., the first item from a task).

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 the VJET Level II score reporting, fluency is taken as a component of oral proficiency. 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 the ease of the underlying encoding process.

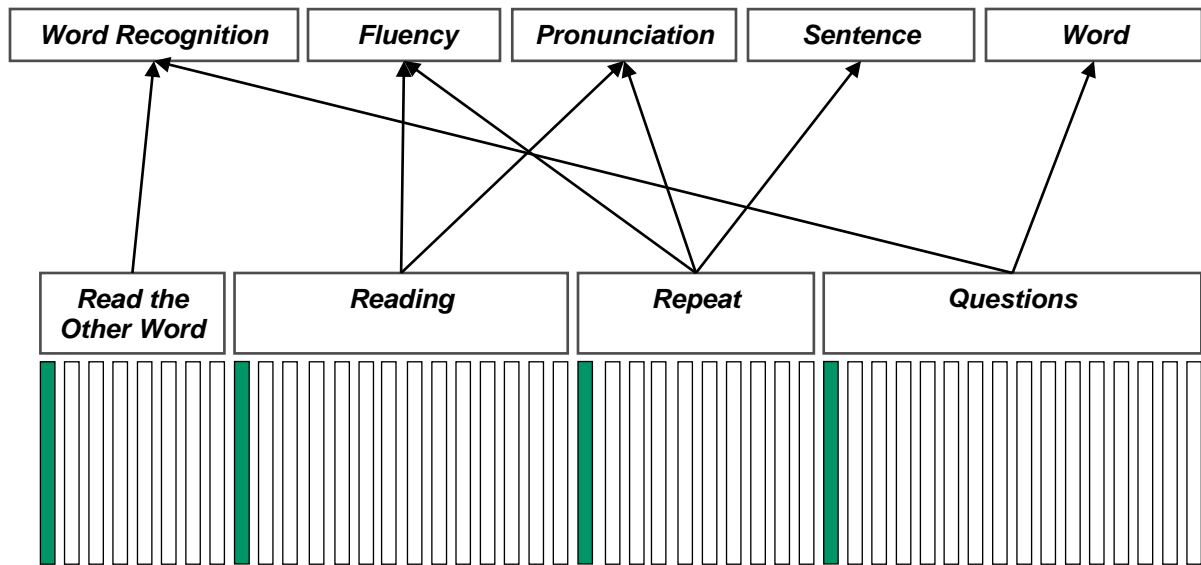


Figure 2. Relation of subscores to item types for the VJET Level II.

The five subscores that make up the Overall score 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). This distinction corresponds roughly to Carroll's (1961) description of a knowledge aspect of language performance and a control aspect. Word Recognition, Word Meaning, and Sentence Structure are associated with content, and Fluency and Pronunciation are associated with manner. The content accuracy dimension counts for 60% of the Overall score and indicates whether or not the student understood the prompt and responded with appropriate content. The manner-of-speaking scores count for the remaining 40% of the Overall score, and indicate whether or not the student 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 hinder oral communication; on the other hand, inappropriate word usage and misapplied syntactic structures can also hinder communication. In the VJET Level II scoring logic, content and manner (i.e., accuracy and control) are weighted almost equally because successful communication depends on both. Table 3 presents the weight of each subscore in the calculation of the Overall score.

Table 3. The weight of each subscore in the VJET Level II as it relates to the calculation of the Overall score.

Subscore	Weight
Content	
Sentence Structure	0.25
Word Recognition	0.10
Word Meaning	0.25
Manner	
Fluency	0.20
Pronunciation	0.20
Total	1.00

In each section of the test, an incoming response is recognized automatically by a Hidden Markov Model (HMM)-based speech recognizer developed from the HMM Tool Kit, (Young, Kershaw, Odell, Ollason, Waltchev, and Woodland, 2000). The acoustic models for the speech recognizer (models of each sound in the language) were trained on data from young non-native speakers of English. In this way, the speech recognizer is optimized for non-native, accented, and age-appropriate speech. The speech recognizer also uses language models to represent the errors and disfluencies that are common for young non-native English speakers for each item. For the content subscores, the system checks for the presence or absence of the correct lexical content in the correct sequence.

The manner-of-speaking subscores are based on speech timing and spectral information. In the Ordinate automated scoring system, words, pauses, syllables, phones, and even some subphonemic events are identified and extracted from the recorded signal for measurement. The manner-of-speaking subscores 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. These base measures are automatically generated and then are recalibrated according to models of expert ratings of pronunciation and fluency. By rescaling the machine scores, the system ensures that the manner-of-speaking subscores optimally predict expert judgments.

The information that forms the basis of each subscore is disjointed from the information underlying the other component subscores. That is, each subscore is based on different, independent aspects of the spoken response material. The content subscores differ in the test section material, while the manner-of-speaking subscores differ in the base measures that are used in the calculation.

4.2 Score Use

Once a test-taker has completed a test, the Ordinate testing system analyzes the spoken performances and posts the scores at www.VersantTest.com within minutes of test completion. Test administrators and score users can then view and print out the test results from www.VersantTest.com, as well as listen to the names of the test-takers in order to verify identity.

Test scores from the VJET Level II can be used to evaluate the level of spoken English skills of individuals entering into, progressing through, and exiting English language courses. Scores can also be used in

evaluating whether a test-taker’s level of spoken English meets a certain established threshold. Versant Benchmarking Kits help score users determine the appropriate threshold for their organization.

Scores may be used for making valid decisions about oral English interaction skills of individuals, provided score users have reliable evidence confirming the identity of the test-takers at the time of test administration. Score users may obtain such evidence either by administering the VJET Level II test themselves or by having trusted third parties administer the test.

Score users have a significant responsibility to ensure the accuracy and thoroughness of interpretation in order to protect the rights of the individual being assessed. It is recommended that other relevant information be considered in combination with test results to enhance overall decisions.

5. Validation

5.1 Validation Study Design

Two indicators of a test’s quality are the test’s reliability and the validity of the interpretations derived from the scores (Alderson, Clapham and Wall, 1995). The statistical data presented in this section provide evidence of the reliability and validity of the VJET Level II.

Responses and test scores from a randomly selected sample of participants were analyzed for the reliability and validity studies. Three groups were organized for these purposes: 1) a non-native learner group (N=166), 2) a group of native children (N=73), and 3) a mixed group (N=175). The mixed group included all the non-native children in the non-native learner group with an additional 5% of native English-speaking children from the native children group to represent performances at the high end of the scale. Table 4 presents the demographics of these three groups of participants. None of the responses from the participants in Table 4 were used to create the models for the automatic scoring or scaling of the test. Therefore, once the models have been created from another sample, they can fairly score the mixed group responses.

Table 4. Demographics of test-takers whose responses and test scores were analyzed in the reliability and validity studies.

Group	N	Male	Female	Age Range	Average Age
1. Non-native	166	78 (47%)	88 (53%)	11 to 16	13.4
2. Native	73	36 (49%)	37 (51%)	11 to 15	12.8
3. Mixed (166 NN, 9 Native)	175	82 (47%)	93 (53%)	10 to 16	13.4

5.2 Reliability

The reliability of a test refers to the precision, consistency and stability of test scores. One estimate of score consistency is the standard error of measurement (SEM). The SEM provides an estimate of the amount of error, due to unreliability, in an individual’s observed test score. Luoma (2004) states, “The SEM shows how far it is worth taking the reported score at face value” (p.183). From an analysis of the native and non-native mixed sample, the SEM of the VJET Level II on the Overall score is 8.7.

Another estimate of score consistency is split-half reliability. Split-half reliability is calculated from data on a single administration of the test. A single test is divided into two equivalent subtests (e.g., a subtest with all the odd numbered items and a subtest with all the even numbered items) and scores from those equivalent halves are correlated. Bachman (2004) states, “The reliability of a test is partly a function of its length, in terms of items it has, so that long tests can generally be expected to be more reliable than short tests” (p.162). For this reason, the correlation is corrected for split-half underestimation by using the Spearman-Brown Prophecy Formula. Reliability values range from 0 (no consistency) to 1 (perfect consistency). Table 5 presents the split-half reliability of the Overall score and subscores of the VJET Level I, using data from the Mixed sample described in Table 4.

Table 5. Split-half reliability of the Overall score and subscores (N=175).

Score	Split-half Reliability
Sentence Structure	0.91
Word Recognition	0.63
Word Meaning	0.63
Fluency	0.91
Pronunciation	0.89
Overall	0.93

The reliability coefficient for the Overall score is greater than those for the individual subscores because the Overall score represents the test-taker’s performance on a broader set of measures than a single subscore. The higher the reliability coefficient, the greater confidence one can place on the consistency of the scores. The high Overall reliability of the VJET Level II is a good indicator that the computerized assessment will be consistent for the same test-taker assuming no changes in the test-taker’s language proficiency level between test administrations. The reliability of Word Recognition and Word Meaning is lower than the other subscores because test-takers’ variability in vocabulary knowledge across test items is relatively large in relation to the sample of vocabulary presented in the test items and used as a basis for the Word Recognition and Word Meaning subscores.

5.3 Validity

Test validity is the extent to which a test measures what it is intended to measure. The degree of validity depends on the evidence supporting the interpretation of test scores. Three lines of such evidence are presented below: content validity, construct validity, and a demonstration of machine scoring accuracy.

5.3.1 Content Validity

Content validity is the degree to which the test items represent the content that the test is designed to measure. Content-related evidence of validity was provided by both expert judgment and empirical item analysis. As described in the Content Development section above, each item was reviewed by subject matter experts to ensure content relevancy and conformity to colloquial usage. In addition, each item was analyzed statistically using data from both native and non-native speakers of English. If an item was not answered correctly by 90% of a sample of natives, then it was unclear if the item was measuring spoken English or another underlying ability, and the item was thus not included in the test. Other

statistical analyses were performed to ensure that each item is effective at discriminating language ability. For more information, see section 3 on Content Development.

5.3.2 Construct Validity

Construct validity is the extent to which test scores can be interpreted as a measure of the intended construct. The test construct for the VJET Level II is facility with spoken English. An indicator of construct validity is the separation of native and non-native speakers of English with regard to Overall scores. The assumption is that native English speakers as a group possess a high degree of facility in spoken English. Therefore, if native English speakers obtain high scores while non-native English speakers are distributed over a wide range of scores, then this expected distinction between the groups lends support to the test's validity.

Figure 3 presents cumulative distributions of Overall scores for the native and non-native samples listed in Table 4.

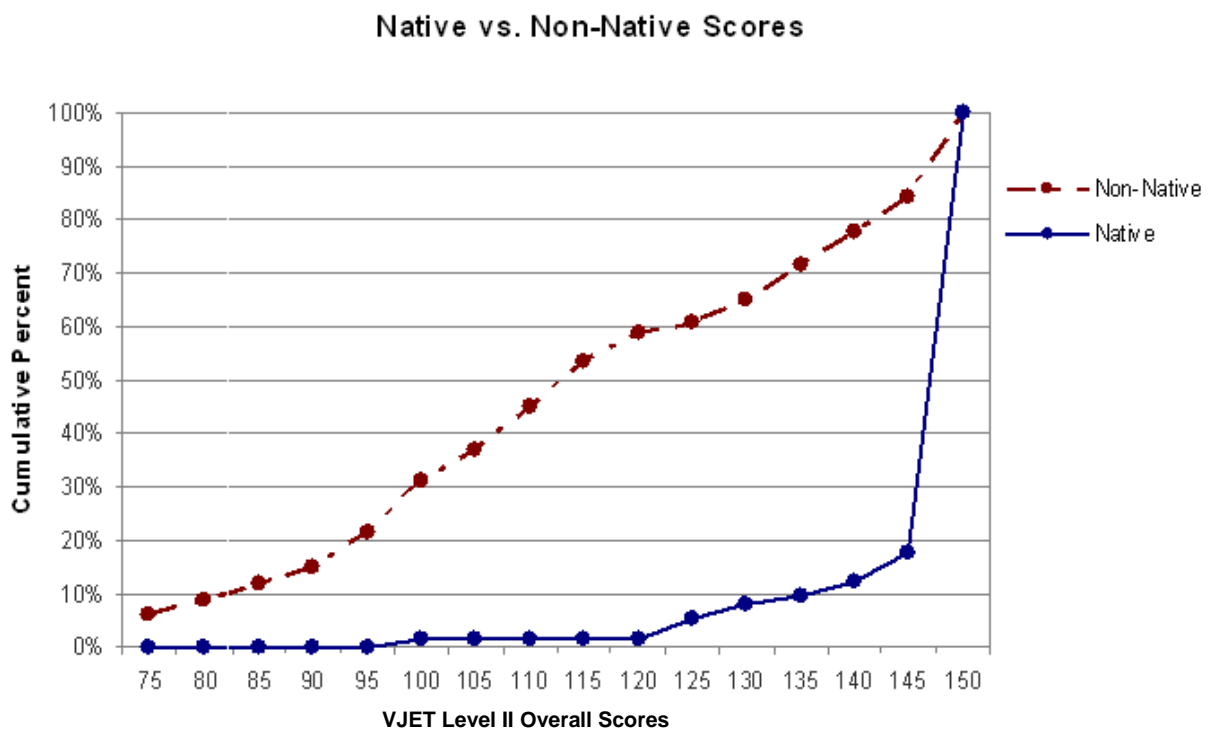


Figure 3. Cumulative distribution functions for native and non-native speakers of English. (Natives N=73, M=148, and Non-Natives N=166, M=118)

The comparison of score distributions shows that native speakers are clustered at the high end of the score scale, whereas English language learners are distributed across a wide range of scores. The mean score for native speakers is 148 while the mean score for non-native speakers is 118. The result of the native/non-native comparisons provides a piece of evidence of the construct validity of the test.

5.3.3 Accuracy of Machine Scores

The VJET Level II is different from most other assessments because of its use of technology to automatically score spoken performances. As evidence of the accuracy of Ordinate's automated scoring, a study was conducted to compare machine-generated scores and scores generated from expert human raters and transcribers on the same items in the test. A high correlation between the two scores suggests that the machine scoring is similar to scores produced by experts.

Using the Mixed sample, responses were transcribed by trained human transcribers and rated independently by expert human raters for pronunciation and fluency. All expert human raters had master's degrees in either TESOL, linguistics or applied linguistics and had several years of experience teaching English to non-native speakers. Using human transcriptions and human ratings, the responses were analyzed and scored without automatic speech processing technologies. These human-generated scores were then compared with the machine-generated scores to determine the accuracy of the automatic scoring. Table 6 below presents correlations between the human-generated and machine-generated scores.

Table 6. Correlation of human-generated and machine-generated scores (N=175).

Score	Correlation
Sentence Structure	0.91
Word Recognition	0.70
Word Meaning	0.90
Fluency	0.76
Pronunciation	0.82
Overall	0.93

Figure 4 shows scatterplots of the Overall human-generated scores and machine-generated scores. In the figure, the data points are centered around the regression line, showing a linear trend. Thus, the scores generated by machine and by professional transcribers and expert raters are closely aligned. In addition, as in Table 6, a strong correlation was found for each of the subscores. The correlations and scatterplot presented suggest that the machine-generated scores for the VJET Level II systematically correspond with human-generated scores, providing evidence that the automatically generated scores are virtually the same as those of expert human raters.

Human-Generated vs. Machine-Generated Overall Scores

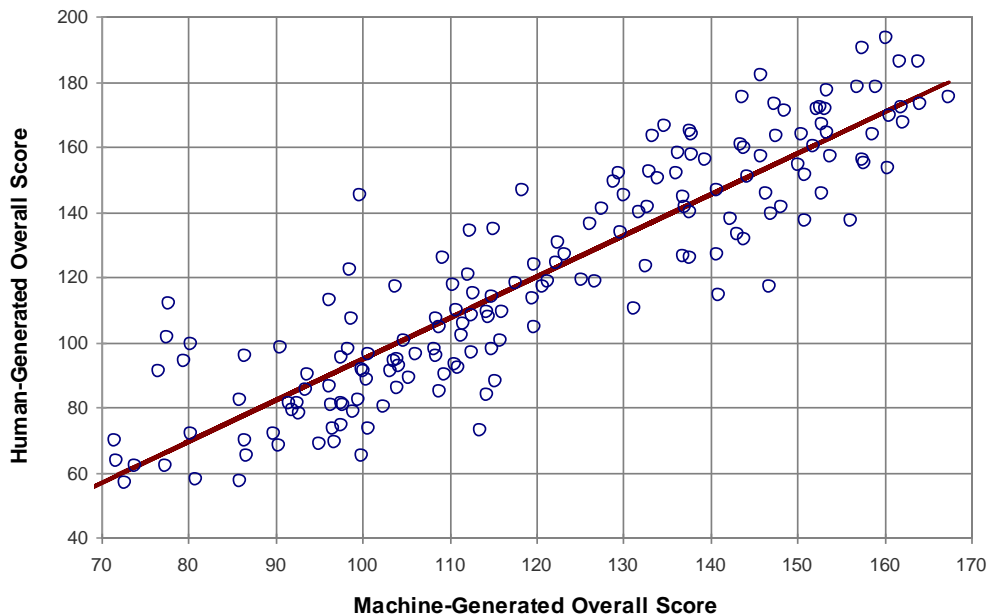


Figure 4. VJET Level II Human-Generated Overall Scores vs. Machine-Generated Overall Scores (n=175, r=0.93).

6. Conclusions

The VJET Level II is an English language test designed to measure the ability of young English language learners to understand spoken English on everyday topics and to respond intelligibly and appropriately at a conversational pace. The test is delivered automatically over the telephone or on a computer and takes about 15 minutes to complete. During the test, the Ordinate testing system presents a series of spoken prompts in English from a variety of speakers and elicits oral responses in English.

A few minutes after the test has been completed, the system analyzes the test-taker's responses and posts scores to www.VersantTest.com. The score report consists of an Overall score and five subscores: Sentence Structure, Word Recognition, Word Meaning, Fluency, and Pronunciation. Together, these scores describe the test-taker's facility in spoken English. Custom cut scores can be established by using Versant Benchmarking Kits.

Reliability of the Overall score is 0.93, suggesting that test scores are internally consistent. Evidence of validity is provided from many sources including expert review of the items, criteria for selecting only items that discriminate abilities among non-native speakers of English, clear separation of performance of native and non-native speakers, and strong correlations with scores generated from human experts.

In sum, the available evidence indicates that VJET Level II test is both a reliable and valid measure of test-takers' facility in spoken English.

7. About the Company

Ordinate Testing Technology: The Ordinate patented 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 Junior English Test is the result of years of research in speech recognition, statistical modeling, linguistics, and testing theory. Ordinate 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, spoken Spanish, spoken aviation English, children's reading assessment, adult literacy assessment, and collections 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 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.


8. References

- Alderson, J.C., Clapham & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.
- Bachman, L.F. (2004). *Statistical analysis for language assessment*. Cambridge: Cambridge University Press.
- Bull, M. & Aylett, M. (1998). An analysis of the timing of turn-taking in a corpus of goal-oriented dialogue. In R.H. Mannell & J. Robert-Ribes (Eds.), *Proceedings of the 5th International Conference on Spoken Language Processing*. Canberra, Australia: Australian Speech Science and Technology Association.
- Butler, F.A., Bailey, A.L., Stevens, R., Huang, B., Lord, C. (2004). Academic English in fifth-grade mathematics, science, and social studies textbooks. *Center for the Study of Evaluation (CSE) Report 642*.
- Carroll, J.B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. *Testing*. Washington, DC: Center for Applied Linguistics.

- Chamot, A.U. & O'Malley, J.M. (1994). *The CALLA handbook: Implementing the cognitive academic language learning approach*. Reading, MA: Addison-Wesley Publishing Company.
- Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), 213-238. Coxhead's website: <http://www.vuw.ac.nz/lals/research/awl/index.html>.
- 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.
- Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.
- Miller, G.A. & Isard, S. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior*, 2, 217-228.
- Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement*, 14, 3-19.
- Standards for Educational and Psychological Testing*. (1999). Washington, DC: American Psychological Association.
- Young, S., D. Kershaw, J. Odell, D. Ollason, V. Valtchev and P. Woodland. 2000. *The HTK Book Version 3.0*. Cambridge, England: Cambridge University.

9. Appendix: Test Materials

Instructions and general introduction to test procedures. Note: These instructions are available in English, Chinese, Japanese, Korean, and Spanish.



TEST INSTRUCTIONS : Versant Junior English Test - Level II

PARTS	INSTRUCTIONS
Before the Test	<ol style="list-style-type: none"> 1. Read these instructions and review the test paper. Make sure that you understand everything. If you don't understand what to do, feel free to ask your teachers or classmates. When you are ready to start, make sure that you have the test paper in hand. 2. ADMINISTRATION <ul style="list-style-type: none"> • <u>Computer Administration</u>: Put on your headset and enter your TIN in the box on your computer screen. • <u>Telephone Administration</u>: Call the telephone number printed on the test paper. When the system asks for your Test Identification Number (TIN), enter it on the telephone keypad. 3. An examiner's voice will guide you through each part of the test. Answer each question in a clear and loud voice. 4. If you don't know the answer, just say "I don't know." in English.
During the Test	<ol style="list-style-type: none"> A. Read the other word. (8 questions) For each question, you will see three words. A voice will read two of the words. You will answer by reading out loud the word the voice did NOT say. B. Reading Sentences. (10 questions) The examiner's voice will ask you to read the sentences out loud, one at a time. Read one sentence out loud and then wait for the next instruction. C. Repeat Sentences. (14 questions) You will hear sentences, one at a time. After you hear each sentence, repeat it exactly as you heard it. Repeat as much of the sentence as you can. D. Short-Answer Questions. (16 questions) For each question, you will see three answers. The examiner's voice will ask you a question. Listen to the question and then read aloud the word that best answers the question. <p>ENDING:</p> <ul style="list-style-type: none"> • <u>Computer Administration</u>: At the end of Part D, you will see "This is the end of the Versant Junior English Test." Click "End Test" on the bottom of the computer screen to end the test. • <u>Telephone Administration</u>: At the end of Part D, you will hear "This is the end of the test. Thank you for calling and goodbye." Then, the test is over. You may hang up.
Tips and Advice	<ul style="list-style-type: none"> • Find a quiet room. Relax and concentrate on the test. Don't take the test around other activities or people that could distract you. • Speak clearly and loudly into the telephone or microphone. Don't be shy! • Give smooth, quick answers. • Say your answers in English. Don't speak Korean or any other language! • If you cannot say the entire answer, say as much as you can. Just do the best you can!

PEARSON

© 2009 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).

Versant Junior English Test - Level II (Demo)

Part C: Repeat.

Please repeat each sentence that you hear.

a voice says, "They told me that yesterday."

and you say, "They told me that yesterday."

Part D: Questions. Now, please just give a simple answer to the questions.

Example: you see the words:

18. cat

book

chair

and a voice says: "Number 18: Which one is an animal?"

and you say, "cat".

1.	morning	night	day
2.	dog	cherry	letter
3.	coin	whale	bed
4.	piano	plane	instrument
5.	man	husband	woman
6.	John	grandmother	uncle
7.	tongue	ear	taste
8.	log	island	jacket
9.	shark	horn	peas
10.	wise	large	assist
11.	ride	draw	wash
12.	broccoli	skates	liquid
13.	dentist	policeman	teacher
14.	bird	bee	butterfly
15.	telephone	handkerchief	purse
16.	pillow	chair	desk


PEARSON

11111 - 1 - #2

© 2012 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

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 1211L