8

Logic and Validation of Fully Automatic Spoken English Test

Jared Bernstein & Jian Cheng

Introduction

Skillful performance in a human language often involves a composite of elementary skills, such that language skills, cognitive skills, and social skills can be conflated in the judgment of a human listener. A recent, computer-based method (Versant for English) provides an estimate of spoken language skills that is relatively independent of the speaker's other social and cognitive skills. The Versant system has been implemented using interactive voice response to administer a standardized spoken English test that measures speaking and listening during a 12-minute telephone call. The system calculates scores on five performance subscales from a set of more basic measures that are produced by automatic speech recognition of examinee responses. Item response theory is used to analyze and scale aspects of examinee performance. The scores are also related to performance rubrics used in criterion-based human scoring of similar responses. This chapter outlines the test construct and describes the scaling methods and validation process with reference to experimental procedures used in developing the test.

Background

Computational and applied linguists have developed various methods for evaluating the performance of different elements and agents in speech communication. These elements include the spoken message source, the transmission medium or system, and the spoken message receiver. The spoken message source and receiver are most often human beings, but they can be speech synthesizers or speech recognizers. The testing methods used to evaluate message sources and message receivers differ by tradition (e.g., applied linguistics or spoken language engineering) and depend also on the population being tested (e.g., automatic systems, or second language learners, or deaf children).

This chapter presents an automatic method for evaluating the spoken language skills of second language learners. In particular, we describe a proficiency test called *Versant for English*, developed by Ordinate Corporation. This test, originally called the *PhonePass* test or the *SET-10*, measures a person's facility in spoken English.

A proficiency test, as such, assumes a domain of knowledge or skill that is independent of any particular instructional curriculum but that measures a construct. The construct of a test is a hypothesis about the attribute (or trait) that the test is designed to measure, for example, "mechanical aptitude" or "language proficiency." The construct of a test is important because it indicates what the test scores should mean; that is, what inference(s) can be drawn from the test scores. The validation of a test is the compilation of evidence that the test is reliable and that the scores do, in fact, reflect the intended construct and do not reflect construct-irrelevant characteristics of the candidates or of the test itself.

Spoken language proficiency tests

Tests of reading, grammar, or vocabulary can be administered quite efficiently. However, traditional speaking/listening tests have been administered by skilled examiners who usually interact with the test-takers one on one, or, at best, listen to tapes one by one. If the assessment of speaking/listening can be automated or even partially automated, then these skills can be tested more often and more reliably across time and place. The end result will be a more accurate and timely evaluation of examinees.

Over the past several decades, human performance in spoken language has traditionally been measured in an oral proficiency interview (OPI) that is judged by the interviewer and, often, by a second human rater. Starting in the 1960's, efforts began to define a construct that would satisfactorily represent general proficiency with the spoken forms of language. Wilds (1975) and Sollenberger (1978) describe the development and use of oral proficiency interviews (OPIs), which were designed by the U.S. Government to measure an examinee's production and comprehension of spoken language during participation in a structured interview with trained interlocutor/raters. The oral proficiency construct was analyzed into a set of level descriptors within each of five subskill areas: comprehension, fluency, vocabulary, grammar, and pronunciation. In the 1980's and 1990's, this general method of interview and level descriptions was adapted and refined for use by other bodies, including the American Council on the Teaching of Foreign Languages (ACTFL) and the Council of Europe. The oral proficiency interview has also been taken as an important validating criterion measure in the development of "indirect" standardized

tests that intend to encompass an oral component, for example the Test of English as a Foreign Language (TOEFL), TSE, and TOEIC tests from the Educational Testing Service. Thus, all these tests rely, at least partly or indirectly, on the OPI oral proficiency construct.

Since the mid-1980's, several experiments have shown that the pronunciation quality of spoken materials can be estimated from direct acoustic measurement of select phenomena in recorded speech samples. Early studies by Molholt and Pressler (1986), by Major (1986), and by Levitt (1991) supported the idea that particular acoustic events in non-native speech could be used to order sets of speech samples by pronunciation. Bernstein et al. (1990) demonstrated that some aspects of pronunciation can be scored reliably by completely automatic methods (see also Neumeyer et al. 1996). These measures of pronunciation quality have some further predictive power, because, in a population of non-natives, the pronunciation of a sample of speakers is a good predictor of overall oral proficiency (Bejar, 1985). The development of Versant for English is an attempt to go beyond this convenient, predictive relation of pronunciation to proficiency and attempt to define automatic procedures that offer a more convincing measurement of the several performance elements that comprise speaking skill.

The remainder of the paper describes an automatically scored 10-minute spoken language test that is delivered by the Ordinate's testing system. First, the target construct of the 12-minute Spoken English Test (Versant for English) is described. Next, the test structure is described in relation to an underlying psycholinguistic theory of speaking and listening. Finally, evidence is presented to establish the valid use of this test as a measure of speaking and listening.

The facility construct

Facility in spoken English

Versant for English was designed to measure "facility in spoken English." We define facility in spoken English to be the ability to understand spoken language and respond intelligibly at a conversational pace on everyday topics. Assuming normal intelligence and basic social skills, this facility should be closely related to successful participation in native-paced discussions – i.e. the ability to track what's being said, extract meaning in real time, and then formulate and produce relevant responses at a native conversational pace. The Versant test measures both listening and speaking skills, emphasizing the candidate's facility (ease, accuracy, fluency, latency) in responding to material constructed from common conversational vocabulary. The test focuses on core linguistic structures and basic psycholinguistic processes.

In contrast to OPI oral proficiency

The Versant construct "facility in spoken English" does not extend to social skills, higher cognitive function, or world knowledge. Nor is the Versant for English test intended to differentiate between examinees' performance on elements that characterize the most advanced range of communicative competence, such as persuasiveness, discourse coherence, or facility with subtle inference and social or cultural nuances. Thus, Versant for English is not a direct test of "oral proficiency" as measured by an oral proficiency interview (OPI), but it shares some key construct elements with such interview tests and will account for much of the true variance

measured by oral proficiency interviews. Because the test measures basic linguistic skills, with emphasis on ease and immediacy of comprehension and production, scores should be appropriate in predicting how fully a candidate will be able to participate in a discussion or other interaction among high-proficiency speakers.

Processing capacity hypothesis

If a test measures only spoken language facility, distinct from other social and cognitive abilities, why should it also be a strong predictor of oral proficiency scores that are designed explicitly to include these other abilities (see the section, "Concurrent Validity with Tests of Related Constructs")? An analogy may be found in the literature on comprehension of synthetic speech. Bernstein and Pisoni (1980) measured students' comprehension of paragraphic material when the paragraphs were read aloud and the students were asked to answer multiple choice questions on the content of the paragraph. The paragraphs were read aloud in two conditions – either by a human talker or by a speech synthesizer. Results suggested that there was no significant decrement in comprehension when students listened to synthetic speech relative to natural human speech. In a later experiment Luce, Feustel, and Pisoni (1983) ran a parallel study but required students to perform a concurrent memory-load task involving visually presented digits that students were asked to recall later. Results from the second study showed a large and significant decrement in comprehension when students listened to synthetic speech in comparison to natural human speech. The authors hypothesized a processing capacity limit to explain the difference in the two experimental results. With no concurrent task, the listeners used as much cognitive capacity as was needed to comprehend the speech samples, and they could extract the words and meanings in the paragraphs adequately in either the human-read or synthesized rendition. However, with the concurrent digit memory task, the listeners still had enough capacity to understand the human speech but did not have the extra capacity required to de-code the synthetic speech. Thus, their comprehension of the synthetic speech suffered.

We hypothesize a similar processing capacity limit relevant to speaking. When a person has limited English skills, the cognitive resources that might be spent on planning a discourse or attending to subtle aspects of the social situation are instead used to find words and expressions that will convey the basic information to be communicated. As a person's command of a language becomes more complete and automatic, more cognitive capacity will be available to apply to the construction of complex argument and/or to the expression of social nuance.

Similarly, in listening, if a person can immediately and completely understand every word and every sentence spoken, then that person will have time to consider the rhetorical tone and the intellectual and social setting of the material. When a person with limited proficiency in a language listens to a connected discourse (even on a familiar topic), that person spends much more time in reconstructing what has been said; thus, there is less time to consider the finer points of the message. Thus, in both listening and speaking, if a person's control of the language is not automatic and immediate, there will likely be a corresponding decrement in the person's ability to use the language for a full range of communication tasks. This hypothesis is consistent with the findings of Verhoeven (1993) that discourse analysis and rhetorical skills will transfer from one language to another. For this reason, over a

range of language proficiencies from beginner to advanced intermediate levels, automaticity in reception and production of basic linguistic forms is a key construct.

Versant for English Test Structure

General structure

The Versant for English test is an examination of speaking and listening in English that is administered over the telephone by a computer system. Candidates' spoken responses are digitized and judged by a specially modified, automated speech recognition (ASR) system. The test presents the examinee with a set of interactive tasks (e.g., to repeat a sentence or answer a question) that require English oral comprehension and production skills at conversational speeds. The test was designed to be particularly appropriate for screening or placement decisions when large numbers of students or candidates are tested and when the examinees are not conveniently available in a single location. The test is intended for use with adult non-native speakers and incorporates fluency, pronunciation, and alacrity in speaking, reciting, and reading aloud; it also incorporates productive control of common vocabulary and of basic sentence structure in repeating sentences and answering short questions. The adult Versant test has been validated with populations as young as 15 years, and newer "junior" forms of the test have been developed for school children.

The Versant for English test, schematized in Table 1, has five parts: Readings, Repeats, Short-Answers, Sentence-Builds, and Open Questions. The first four parts (A-D) are scored automatically by machine, while the fifth part (E) collects three 20-second samples of the examinee's speech that can be reviewed by score users. General instructions are printed on the back of the test paper, and specific instructions for the five parts of the test are spoken by the examiner voice and printed verbatim on the face of the test sheet. Items are presented in various item voices that are distinct from the examiner voice that introduces the sections and provides instructions.

Table 1. Versant for English Test Design

Part	Item Type	Target Skills	Item Count
A	Read aloud	Basic listening, reading fluency, pronunciation	8
B	Repeat sentence	Listening, vocabulary, syntax, fluency	16
C	Short answer	Vocabulary in syntactic context	24
D	Sentence build	Listening, vocabulary, syntax, fluency	10
E	Open response	Discourse, fluency, pronunciation, vocabulary	3

Versant tasks are designed to be simple and intuitive both for native speakers and for proficient non-native speakers of English. Items cover a broad range of skill levels and skill profiles. They are designed to elicit examinee responses that can be analyzed by machine to produce measures that underlie facility with English, including fluency, listening, vocabulary, sentence mastery, and pronunciation.

Item design specifications

All item material was crafted specifically for the test, but it follows lexical and stylistic patterns found in actual conversation. The items themselves are recorded utterances

that are presented in a specified task context. To ensure conversational content, all materials use vocabulary that is actually found in the spontaneous conversations of North Americans.

Vocabulary: Versant for English vocabulary is taken from a list of 7,727 word forms that occurred more than 8 times in the Switchboard corpus, a 3-million word corpus of spontaneous American conversation) available from the Linguistic Data Consortium (http://www.ldc.upenn.edu). Items may include any regular inflectional forms of the word; thus, if "folded" is on the word list, then "fold," "folder," "folding," and "folds" may be used.

Voices: The audio item prompts are spoken by a diverse sample of educated native speakers of North American English. These voices are clearly distinct from the examiner voice that announces the general instructions and the task-specific instructions.

Speaking style: The Repeat and Short Question items are written in a non-localized but colloquial style, with contractions where appropriate. Thus, a prompt will be written out (for the item speaker to recite) in a form such as, "They're with the contractor who's late." rather than "They are with the contractor who is late." The people who speak the items are instructed to recite the material in a smooth and natural way; however, normally occurring variation is permitted in speaking rate and pronunciation clarity between speakers and items.

World knowledge and cognitive load: Candidates should not need specialized world knowledge or familiarity with local cultural referents to answer items correctly. Versant for English items are intended to be within the realm of familiarity of both a typical North American adolescent and an educated adult who has never lived in an English-speaking country. The cognitive requirement to answer an item correctly should be limited to simple manipulations of time and number. Operationally, the cognitive limit is enforced by requiring that 90% of a norming group of native speakers can answer each item correctly within six seconds. Versant for English items should not require unusual insight or feats of memory.

Psycholinguistic Performance Theory

Psycholinguistic research has provided evidence for the operation of internal processes that are used when people speak and listen. Some of these processes have a parallel in linguistic theory while others do not. Adapting from the model proposed by W. J. M. Levelt in his book *Speaking* (1988), we can posit the psycholinguistic processing steps shown in Figure 1. A speaker encodes declarative, social and discourse information into spoken utterances, and the listener needs to decode this information with reference to a common set of language structures. To understand what is said in a conversation, a listener needs to hear the utterance, extract lexical items, identify the phrase structures manifest in the words, decode the propositions

carried by those phrases in context, and infer from them the implicit or explicit demands. When speaking, a person has to perform a similar set of operations but in approximately the reverse order. Note that the experimental evidence is equivocal about the exact order (or interleaving) of these operations and their modularity. However, all these operations must be accomplished in order to speak and understand speech. During a conversation, every active participant is performing either the understanding processes or the speaking processes or some of both.

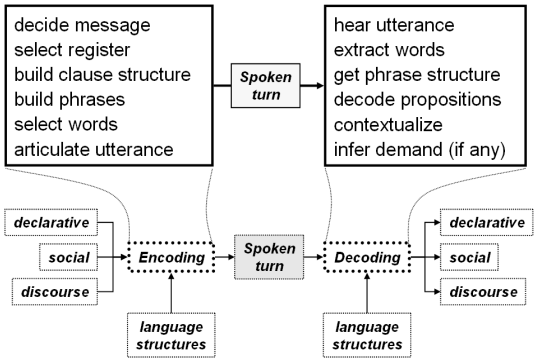

Figure 1. Internal processing flow in a model of speaking and listening.

If a test measures a candidate's facility in grasping spoken utterances in real time and producing relevant, intelligible responses at a native conversational pace, then it should be covering the basic components of psycholinguistic processing. Because the task items in the Versant for English test present de-contextualized voice recordings to the candidate, each of which elicits a spoken response, each task item exercises the processing elements shown in Figure 1 except for register selection and the contextualization.

Versant for English is a direct test of facility with spoken English materials. Performance on the Versant tasks provides evidence of ability to participate in English conversation to the extent that those tasks require many of the same skills used in natural conversation. Following the theoretical framework of Levelt (1988), the corresponding skills include the following:

Generation/speaking elements	**Sampled in Versant items**
1. decide message	all items
2. select register	—
3. build clause structure	long repeats, questions, sentence builds
4. build phrases	questions
5. select lexical items	repeats, questions, sentence builds
6. articulate utterance	all items
Listening/understanding elements	**Sampled in Versant items**
1. hear utterance	all items
2. recognize lexical forms	repeats, questions, sentence builds
3. extract linguistic structure	long repeats, questions, sentence builds
4. decode propositions	questions
5. contextualize	—
6. infer demand	all items

It can be seen from this list that the Versant for English items exercise all the processing steps in listening and speaking except register selection and contextualization. The ability to select an appropriate register and to accommodate material for current context could be tapped in Versant for English, but these functions would require more elaborate items than are now included in the test.

Scoring

General approach to scoring

The Versant for English Score Report gives an overall score and four component subscores. The Overall score represents a measure of the examinee's facility in spoken English. It is calculated as a weighted average of the four subscores, as follows:

- **30%: Sentence Mastery –** understand, recall, and produce phrases and clauses in complete sentences
- **20%: Vocabulary** – understand words spoken in sentence context
- **30%: Fluency –** rhythm, phrasing and timing, as evident in constructing, reading and repeating sentences
- **20%: Pronunciation –** intelligible consonants, vowels, and lexical stress

These four subscores are relatively independent of each other, although some of them represent different measures derived from the same responses, as suggested above in the aligned list of process elements and item types. For example, the Pronunciation subscore is based on responses to the Reading, the Repeat, and sentence build sections.

The subscores are of two logical types, categorical and continuous. The first type, categorical, is based on the correctness of the content of the response. That is, the subscore reflects the number of found in the exact words spoken in the responses. This type of score shows how well the candidate understood the item and provided a complete and correct response to it. Scores of this type (Sentence Mastery and Vocabulary) comprise 50% of the Overall score. The second type of score, continuous, is based on the manner in which the response is spoken (pronunciation and fluency); these scores comprise the remaining 50% of the Overall score.

Criterion scoring by human listeners

The continuous subscores (pronunciation and fluency) were developed with reference to human judgments of fluency and pronunciation. The rubrics for these criterion scores and the level descriptions of the skill components were developed by expert linguists who are active in teaching and evaluating spoken English.

Three master graders were asked to develop, apply, and refine the definitions of two scoring rubrics: fluency and pronunciation. The rubrics include definitions of these skills at six levels of performance as well as criteria for assigning a "no response" grade. The master graders scored a large, random sample of examinee

responses and tutored other human graders in the logic and methods used in the criterion grading.

Human graders assigned over 129,000 scores to many thousands of responses from hundreds of different examinees. Item response analysis of the human grader scores indicates that human graders produce consistent fluency and pronunciation scores for the Versant materials, with single-rater reliabilities between 0.82 and 0.93 for the various subskills.

Machine scoring

All Versant for English reported scores are calculated automatically using speech recognition technology. ASR in Versant for English is performed by an HMM-based speech recognizer built with Cambridge University Engineer Department's HTK toolkit. The acoustic models, pronunciation dictionaries, and expected-response networks were developed at Ordinate Corporation using data collected during administration of Versant for English.

The acoustic models consist of tri-state triphone models using seven Gaussian mixtures that specify the likelihood of 26-element cepstral feature vectors. These models were trained on a mix of native and non-native speakers of English using speech files collected during administration of Versant for English. The expected response networks were formed from observed responses to each item over a set of over 370 native speakers and 2,700 non-native speakers of English. The speech data from one quarter of the speakers was reserved for testing only.

As outlined above, subscores are calculated by two main techniques: analysis of correct/incorrect responses and function approximation using statistical output from the speech recognizer.

First, each utterance is recognized and categorized. In the Repeat and Sentence Build section of Versant for English, the accuracy of the response can be determined by reference to the number of words inserted, deleted, or substituted by the candidate. These item-level scores are then combined to give a "Sentence Mastery" component measure. This combination is done using Item Response Theory (IRT) such that both the difficulty of the item and the expected recognition performance of the item contribute to its weight. For example, very difficult items will have a small effect on the measure of a low-level examinee but a larger effect on more proficient examinees. Similarly, items that are often misrecognized will have lower weight. Using the same method, a correct/incorrect decision for each item in Parts C (see Table 1) contributes to the "Vocabulary" component measure. These correct/incorrect decisions are based partly on observed responses to the item by native and non-native speakers.

In the Reading, Repeat, and Sentence Build parts of the test, the responses consist of a complete phrase or sentence. Thus, in addition to the accuracy of the response, we can also make use of the alignment and other properties of the speech signal to further judge the speaker's ability. Signal analysis routines perform a set of acoustic base measures on the linguistic units (segments, syllables, and words) and return these base measures.

Different base measures are combined in different ways into the two continuous measures – Pronunciation and Fluency. The combination is achieved by a parametric function optimized against judgments from human raters on these same criteria. The goal of the function is that for each examinee, the expected difference between the

human-judged ability and the component measure should be minimized. An overall summary grade, representing facility in spoken English, is calculated as a weighted combination of the continuous measures and the categorical measures.

Evidence of Validity

An assertion that scores from a given test are valid for a particular use can be supported by many kinds of evidence. We have gathered seven kinds of evidence for the assertion that the Versant for English instrument provides a valid measure of facility in spoken English:

1. Test material samples key aspects of the performance domain.
2. Human listeners can estimate candidate skills reliably from the recorded responses.
3. Machine subscores and the Overall score are reliable.
4. Uniform candidate groups get similar scores.
5. Different subscores are reasonably distinct from each other.
6. Machine scores correspond to criterion human judgments.
7. Scores correlate reasonably with concurrent scores from tests of related constructs.

Test material samples the performance domain

As outlined in the sections above, the items of Versant for English are designed to conform to the vocabulary and register of colloquial American English. The items present a quasi-random sample of phrase structures and phonological styles that occur in spontaneous conversation, while the vocabulary is restricted to the high-usage sector of the vernacular lexicon. In particular, the requirement that at least 90% of educated adult native speakers perform correctly on every item suggests that the tasks are within reasonable expectations for all performance elements, including underlying abilities like memory span. The Reading, Repeat, Short Question, and Sentence Build items offer an opportunity for candidates to demonstrate their English skills in integrated performances that exercise most basic psycholinguistic components of speaking and listening.

Human listeners estimate candidate skills reliably

As introduced above (see "Criterion Scoring by Human Listeners"), human listeners judged over 129,000 individual item responses and produced orthographic transcriptions of 247,000 responses during the original development and validation of the Versant for English test. Applying item response theoretic analyses (Wright & Stone, 1979) to these human judgments and transcriptions, we see from the reliability data that human listeners do make consistent judgments of the elemental abilities represented in the Versant for English subscores (see Table 2).

We sampled a set of 50 speakers whose test responses were completely transcribed by human listeners and whose responses had human ratings of fluency and pronunciation. We used these human-generated data to derive scores that are parallel to the machine-generated scores of Versant for English. These ratings were reduced to ability subscores for each individual using a single-dimensional IRT

analysis with a Rasch model. In addition, each individual's responses to the Vocabulary, Repeat, and Sentence Build items on the test were transcribed by a human listener, and the number of word errors was calculated. These results were also analyzed using IRT to give a "human-based" ability in vocabulary and sentence mastery. Finally, the four scores for each individual were combined (using the same linear weighting as for the Versant Overall facility score) to give a "human" Overall grade for each individual.

Reliability of the human subscores is in the range of that reported for other human-rated language tests, while the reliability of the combined human Overall score is greater than that normally found for most human-rated tests. These results support the presumption that candidate responses to the items in a single 10-minute test administration provide an adequate sample of spoken data upon which to base meaningful and reliable skill estimates.

Machine scores are reliable

It is not too surprising that a machine will score a single item response consistently, but we want to know whether the Versant system will score many responses from a given candidate in a manner reflecting the relative consistency of those performances. The machine score reliabilities displayed in Table 2 suggest that the speech processing technology used in Versant can transcribe the short utterances elicited from non-native speakers by Versant for English nearly as well as a human listener can transcribe them. Further, the data suggest that the machine's pronunciation and fluency judgments are generally similar in reliability to that observed with a highly trained human listener. Both the human and machine Overall scores show a reliability of 0.94. The reliability is equally high whether the tests are hand transcribed and judged by a human listener or whether the system operates without human intervention.

Table 2. Human and Machine Score Reliability for Four Subscores and the Overall Score (n = 50)

Subscore	One Human Rater	Machine
Sentence Mastery	0.96	0.93
Vocabulary	0.85	0.88
Fluency	0.98	0.95
Pronunciation	0.98	0.97
Overall	0.98	0.97

Uniform candidate groups get similar scores

A common form of test validation is to check whether the test produces expected score distributions for familiar populations with well-understood ability levels in the target construct. One may presume that there are relatively uniform groups of candidates who have very distinct levels of the construct being measured. One would expect that the distribution of scores from a test of proficiency in algebra, for example, would be different for different groups, such as 8-year-old children, high-school students, and professional mathematicians.

Figure 2 show the cumulative density distribution of Versant for English Overall scores for four groups of candidates. The score range displayed on the abscissa extends from 10 through 90, as the Overall scores are calculated inside the Versant system. (Note that scores are reported only in the range from 20 to 80, with scores below 20 Scores below 20 are reported as 20 and scores above 80 are reported as 80.)

The thick black line, rightmost in Figure 2, shows the data for a balanced set of 775 native speakers. The native group consisted of approximately 33% speakers from U.K. and 66% speakers from the USA and had a female/male ratio of 55/45. Ages ranged from 18 to 75. 79% of whom got a score of 80 and only 5% of whom got a score lower than 70.

The thick gray line in Figure 2 displays the cumulative score distribution for a norming group of 606 non-native speakers, ages 12 to 56, balanced in gender, and representing 40 different native languages. Scores from the norming group form a quasi-normal distribution over most of the score range, with a median score of 41.

The two thin lines (black and gray) show two approximately homogeneous populations. The thin black line, leftmost in Figure 2, shows a group of 90 first-year students at a Japanese university. They are all the same age and all studied the same English curriculum for five years; their scores range from 20 to 60. The thin gray line is a group of 170 international graduate students seeking qualification to work as university teaching assistants. Their scores range from 40 to 80.

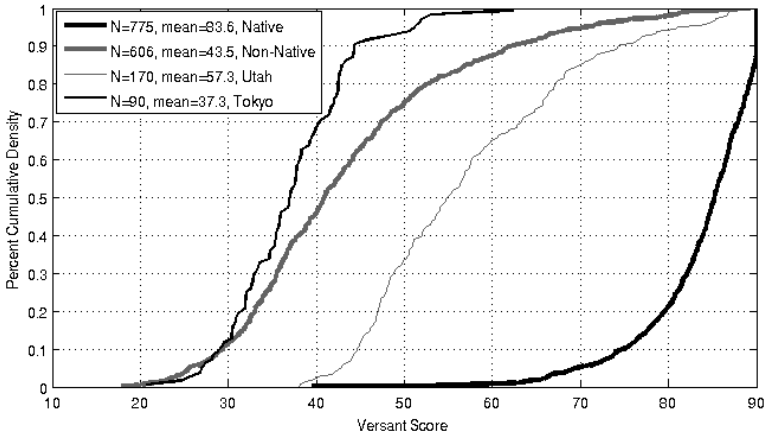

Figure 2. Cumulative distribution of Versant overall score for various populations.

Subscores are reasonably distinct

Versant for English was designed to measure an array of subskills that, when taken together, will provide a reasonable estimate of a more general "facility" in spoken English. The subskill scores are based on distinct aspects of the candidate's performance, and they are identified by names that reflect the performance criteria they are intended to measure. Thus, the Vocabulary score is derived only from items wherein the response task principally involves the immediate recognition and understanding of spoken words and the production of related words. The Fluency score is derived only from measures of a candidate's rhythm, pace, and phrasing while producing complex material, as in Repeat and Sentence Build responses.

Over many candidate populations, various measures of second language abilities will correlate to some degree, often with coefficients in the range 0.5 to 0.9. If subscores correlate too highly, it might indicate that they are two different labels for a common ability. Yet, in some special populations the correlation between certain language skills may be very low, as between reading fluency and repeat accuracy in a group of illiterate native speakers.

Table 3. Correlation Coefficients Between Versant for English Subscores

	Voc.	**SentM**	**Fluency**	**Pron**	**Overall**
Vocabulary	1.000	0.73	0.61	0.65	**0.84**
Sentence Mastery		1.00	0.67	0.71	**0.88**
Fluency			1.00	0.92	**0.90**
Pronunciation					**0.92**

Table 3 shows that the machine subscores correlate with each other, with coefficients in the range 0.61 to 0.92; they correlate with the Overall score in the range 0.84 to 0.92. An interesting case appears in Figure 3, a scatter plot of the Sentence Mastery and Fluency scores for a non-native norming set of 603 candidates. For each candidate, these two scores are measured from the same utterances exactly, but the correlation, as shown in Table 3, is only 0.67.

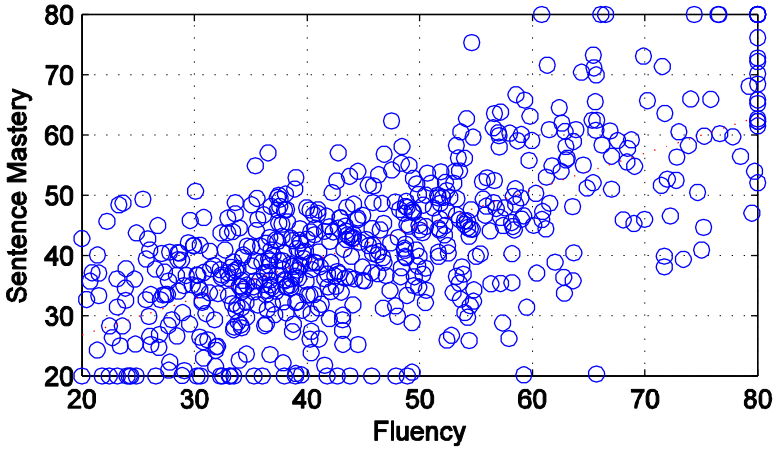

Figure 3. **Sentence Mastery** versus Fluency (n = 603; r = 0.67).

Just as there are candidates who can read aloud but cannot speak fluently, so there are candidates who can repeat a long, complex sentence but cannot do so fluently. There are also candidates who can repeat a short utterance quite fluently, but cannot grasp, understand, and reproduce a longer, more complex sentence. The different subscores reflect these differences.

Machine Scores Correlate with Human Judgments

Scores must be reliable to be valid, but reliability alone will not establish validity. A reliable test score could be a consistent measure of the wrong thing. The human scores for the Overall facility construct exhibit a reliability of 0.98 and it is 0.97 for the machine scores, but we need to know whether the machine scores actually match the human listener scores. Correlations between machine and human subscores in Table 4 show a consistent close correspondence between human judgments and machine scores. The correlation coefficients for the two categorical scores (Vocabulary and Sentence Mastery) are 0.93, 0.94, indicating that the machine recognition algorithms that count for 50% of the score produce measures in close accord with the human-derived scores. Both the algorithmic fluency and pronunciation measures match the human judgments with correlation coefficients of 0.89.

Table 4. Correlations between Machine and Human Scores for Overall Score and Subscores.

Score	Correlation
Overall	0.97
Sentence Mastery	0.93
Vocabulary	0.94
Fluency	0.89
Pronunciation	0.89

We selected a balanced subset of 50 Versant testing candidates whose data had sufficient coverage of human transcriptions and human fluency and pronunciation grades to provide a fair comparison between the human and machine grades for the Overall scores.

Figure 4 shows a scatter plot of the Overall human grades against the Versant Overall grade. The correlation coefficient for this data is 0.97, which compares well with the single-rater reliability we observed for the human-rated Overall score of 0.98. That is, the machine grades agree with the aggregate human judgments about as well as single human raters agree with the aggregate human judgment.

It is interesting to note that the close score correspondence extends over the whole range of scores. Candidates in the 20–40 range have difficulty producing a sentence of four words length, while candidates in the 65–80 range are usually quite fluent and able to generate spoken English at a native or near native pace in paragraphic chunks. The next section presents correlations of Versant scores with other human-graded tests that have somewhat divergent target constructs.

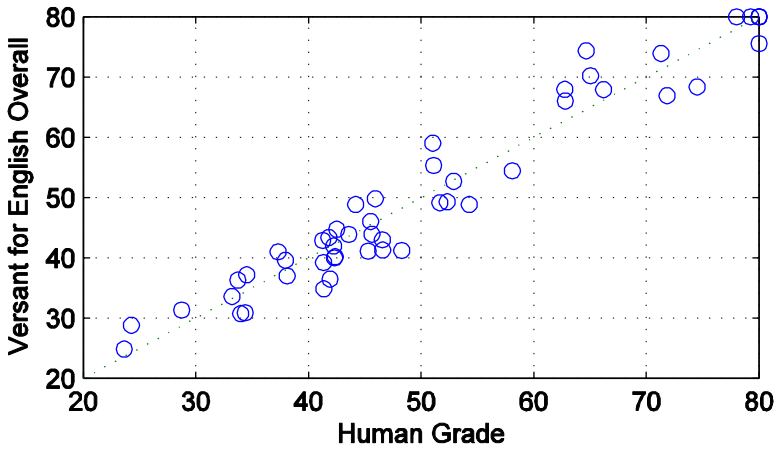

Figure 4. Versant overall facility in spoken English vs. human-rater overall grade (n=50; r = 0.97).

Scores Correlate with Concurrent Scores from Related Tests

The predictive validity of Versant testing as a measure of "oral proficiency" has been studied at several sites. A group of 51 technical visitors to a U.S. Government training program in Texas took oral proficiency interviews (ILR OPI conducted by U.S. government examiners) and also took Versant for English. The correlation between the ILR OPI speaking scores and the Versant Overall scores was 0.75. Because the OPI's inter-rater reliability is about 0.76, a correlation of 0.75 with a two-rater average suggests that Versant Overall scores match the OPI scores about as well as the individual expert human ratings match each other.

We have also examined the correlation between the Versant scoring and scores from other well-established tests of English. These results are shown in Table 5. Cascallar and Bernstein (2000) conducted a study of Versant scoring for a New York State agency. The Versant test was administered concurrently with the Test of Spoken English (TSE) scored at the Educational Testing Service. The subjects were a balanced group of Spanish, Chinese and Russian native speakers.

TSE is delivered in semi-direct format, which maintains reliability and validity while controlling for the subjective variables associated with direct interviewing. The TSE score should be a reflection of an examinee's oral communicative language ability on a scale from 20 to 60 (from "No effective communication" to "Communication almost always effective"). Human raters evaluate speech samples and assign score levels using descriptors of communicative effectiveness related to task/function, coherence and use of cohesive devices, appropriateness of response to audience/situation, and linguistic accuracy.

A raw correlation of 0.88 was measured between the TSE and Versant for English over a sample of subjects relatively evenly spread over the TSE score scale. A corrected validity coefficient of 0.90 was found with respect to TSE as a criterion measure. These results, coupled with the measured reliability of Versant for English, indicate that Versant for English can produce scores that measure a similar

underlying construct to that of TSE. Furthermore, Versant scores can be used to infer TSE scores for the same subject with a mean square error of 5.1 TSE scale points. Since the TSE scale is quantized in 5-point steps, this indicates that a Versant score can predict a subject's TSE score within one score step in most cases.

Table 5. Correlation with Concurrent Scores from Tests with Related Constructs

Test	Correlation with Versant Scores	N
TSE	0.88	59
ILR OPI	0.75	51
TOEFL	0.73	418
TOEIC	0.71	171

The TOEFL correlation shown in Table 5 was calculated from a pool of Versant for English candidates who also reported a TOEFL score. These 418 candidates were repeatedly re-sampled according to the reported distribution of TOEFL scores, to establish a correlation estimate that would be consistent with the range and distribution of TOEFL scores as actually observed worldwide. Because TOEFL is a test of reading comprehension, English structure, and listening (with no speaking component), a correlation like 0.73 may be expected, as would be expected between any two tests of related but different language skills over most populations.

The study of Versant for English in relation to TOEIC was performed in Japan at the Institute for International Business Communication (IIBC). IIBC selected a stratified sample of 171 adults who had recently taken the TOEIC. The sample of candidates fairly represented the range and approximate distribution of TOEIC candidates that are tested in Japan. The correlation of 0.71 between Versant for English and the TOEIC is in the expected range, in that the TOEIC is primarily a reading and listening test, and there is a strong reading component even in the listening section of the test.

Notably, the Versant scoring predicts the scores of the two concurrent tests (TSE and ILR OPI) that are primarily speaking tests and predicts these as well as or better than do the trained, expert raters on those tests. This is true even though the scoring rubrics for both the TSE and the ILR OPI include rhetorical and sociolinguistic aspects of speaking performance that are clearly outside the realm of the scoring algorithms used in Versant for English.

Discussion

The Versant for English test has been designed to exercise and measure the performance of basic psycholinguistic processes that underlie speaking and listening in spontaneous conversation. The test is delivered and scored completely automatically. Yet its scores seem to correspond closely to human judgments of communicative effectiveness as measured in traditional direct and indirect speaking tests.

We have proposed a hypothesis to explain this close correspondence. The hypothesis posits that limits on cognitive processing capacity may explain this extension of the scoring to rhetorical and sociolinguistic properties of speaking performance. It seems that this hypothesis could, in principle, be tested by following

the lead of Luce et. al., and measuring the decrement in discourse cohesion that results when a highly skilled talker is burdened with a difficult collateral task while speaking. Another approach to testing the hypothesis, in principle, would be to measure language independent aspects of discourse cohesion, for example, in two spoken performances on the same topic by the same person in two languages in which the speaker has widely different levels of speaking facility.

In the time since this chapter was drafted and its publication, Ordinate has also built Versant tests for Spanish and Dutch. These tests have been validated with much more thorough comparison to concurrent human-scored interview tests, and the strong relation between machine scores of *facility* and human ratings of interview performance are consistently high.

Finally, it may be noted that the Versant for English test has been available for several years now by telephone for use from anywhere in the world, seven days a week, at any time of the day. Test results are available, on the Ordinate website www.ordinate.com, within about a minute after a test is completed. In operation, the Versant for English scores are more reliable than those from any operational human-scored tests, and thus the Versant testing service may offer a convenient alternative to traditional spoken language testing methods.

References

Bejar, I. (1985). *A Preliminary Study of Raters for the Test of Spoken English*. Research Report RR-85-5, Educational Testing Service, Princeton, NJ.

Bernstein, J., Cohen, M., Murveit, H., Rtischev, D., & Weintraub, M. (1990). Automatic evaluation and training in English pronunciation. In 1990 *International Conference on Spoken Language Processing*, Kobe, Japan: Acoustical Society of Japan, 1 pp. 185-1188.

Bernstein, J. & Pisoni, D. (1980). Unlimited text-to-speech device: Description and evaluation of a micro-processor-based system. *1980 IEEE International Conference Record on Acoustics, Speech, and Signal Processing*, April, 576-579.

Cascallar, E. & Bernstein, J. (2000). *Cultural and functional determinants of language proficiency in objective tests and self-reports*. A paper at the American Association for Applied Linguistics (AAAL-2000) meeting, Vancouver, British Columbia. March, 2000.

Levelt, P. (1988). *Speaking*. Cambridge, Massachusetts, MIT Press.

Levitt, A. (1991). Reiterant speech as a test of non-native speakers' mastery of the timing of French. *J. Acoustical Society of America*. vol. 90 (6), 3008-3018.

Luce, P., Feustel, T. & Pisoni, D. (1983). Capacity demands in short-term memory for synthetic and natural word lists. *Human Factors, 25,* 17-32.

Major, R. (1986). Paragoge and degree of foreign accent in Brazilian English. *Second Language Research* , vol. 2 (1), 53-71.

Molholt, G. & Pressler, A. (1986). Correlation between human and machine ratings of English reading passages. In Stansfield, C. (Ed.), 1986, *Technology and Language Testing*, a collection of papers from the Seventh Annual Language Testing Research Colloquium, held at ETS, Princeton, NJ, April 6-9, 1985, TESOL, Washington D. C.

Neumeyer, L., Franco, H., Weintraub, M. & Price, P. (1996). Automatic Text-independent pronunciation scoring of foreign language student speech. In Bunnell, T. (ed.) *Proceedings ICSLP 96: Fourth International Conference on Spoken Language Processing, 3,* 1457-1460.

Sollenberg, H. (1978). Development and current use of the FSI oral interview test. In Clark, J. (ed.) *Direct testing of speaking proficiency: theory and application*. Educational Testing Service, Princeton, NJ.

Verhoeven, L. (1993). Transfer in bilingual development: The linguistic interdependence hypothesis revisited, *Language Learning, 44,* 381-415.

Wilds, C. (1975). The oral interview test. In Jones, R. & Spolsky, B. (eds.) *Testing Language Proficiency.* Center for Applied Linguistics, Arlington, VA.

Wright, B. & Stone, M. (1979). *Best Test Design*. Chicago, IL: MESA Press.