COMPUTER ESTIMATION OF SPOKEN LANGUAGE SKILLS

Jared Bernstein, Ognjen Todic, Brent Townshend, & Eryk Warren

Ordinate Corporation
1040 Noel Drive
Menlo Park, California
94025 USA

email: jared@ordinate.com

ABSTRACT

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 judgement of a human listener. A new, computer-based method (PhonePass testing) provides an estimate of spoken language skills, relatively independent of the speaker's other social and cognitive skills. The PhonePass system has been implemented using interactive voice response to administer a standardized spoken English test that measures speaking and listening during a 10-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. The test construct is outlined, and the scaling methods and validation process are described with reference to experimental procedures that were used in the development of the SET-10 test, a standardized instrument.

1. BACKGROUND

Computational and applied linguists have developed various methods for evaluating the performance of different elements and agents in speech communication. These 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 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 paper presents an automatic method for evaluating the spoken language skills of second language learners. In particular, we describe a proficiency test called SET-10, developed by Ordinate Corporation, that 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 not construct-irrelevant characteristics of the candidates or of the test itself.

1.1 Spoken Language Proficiency Tests

Tests of reading, or grammar and 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 [13] and Sollenberger [11] 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 ıefined 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 TOEFL, TSE, and TOEIC tests from the Educational Testing Service. Thus, all these tests rely, at least partially or indirectly, on the OPI oral proficiency construct.

Since the mid 1980's, several experiments have shown that pronunciation quality of spoken materials can be estimated from direct acoustic measurement of select phenomena in recorded speech samples. Early studies by Molholt & Pressler [9], by Major [8] and later by Levitt [6] 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. [2] demonstrated that some aspects of pronunciation can be scored reliably by completely automatic methods (see also Neumeyer et al. [10]). 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, [1]). The development of the PhonePass test 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 PhonePass testing system. The target construct of the 10 minute Spoken English Test (SET-10) is described, then the test structure is described in relation to an underlying psycholinguistic theory of speaking and listening, and then we present the evidence that has been marshaled to establish the valid use of this test as a measure of speaking and listening.

1.2 Facility Construct

1.2.1 Facility in Spoken English

The SET-10 test 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 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.

1.2.2 In Contrast to OPI Oral Proficiency

The PhonePass construct "facility in spoken English" does not extend to include social skills, higher cognitive function, or world knowledge. Nor is the PhonePass test intended to differentiate between examinees' performance 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, a PhonePass test 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.

1.2.3 Processing Capacity Hypothesis

If a test only measures spoken language facility, distinct from other social and cognitive abilities, why should it also be a strong predictor (see section 5.7 below) of oral proficiency scores that are designed explicitly to include these other abilities? An analogy may be found in the literature on comprehension of synthetic speech. Bernstein & Pisoni [3] 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 student comprehension when listening to synthetic speech relative to natural human speech. In a later experiment, Luce, Feustel, & Pisoni [7] ran a parallel study but required the subjects to perform a concurrent memory-load task involving visually presented digits that subjects were asked to recall later. Results from the second study showed a large and significant decrement in comprehension when listening 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 samples. 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 otherwise 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 that needs to be communicated. As a person's command of a language becomes more complete and automatic, there will be more cognitive capacity available to apply to the construction of complex argument or to 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), much more time is spent in reconstructing what has been said, therefore the listener does not have as much 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 [12] 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.

2. PHONEPASS SET-10 TEST STRUCTURE

2.1 General Structure

The PhonePass SET-10 test is an examination of speaking and listening in English that is administered over the telephone by a computer system. Candidate's spoken responses are digitized and judged by a specially modified speech recognition system. The test presents the examinee with a set of interactive tasks (e.g. 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, as well as productive control of common vocabulary and basic sentence structure in repeating sentences and answering short questions. The test seems to work well with candidates as young as 12 years, but has not been validated with populations younger than 15 years.

The SET-10 test has five parts: Readings, Repeats, Opposites, Short-Answers, and Open Questions. The first four parts are scored automatically by machine, while the fifth part collects two 30-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: *PhonePass SET-10 test design*

	Item Type	**Target Skills**	**Item Count**
A	Read aloud	Basic listening, reading fluency, pronunciation	8
B	Repeat sentence	Listening, vocabulary, syntax, fluency	16
C	Opposite word	Listening-vocabulary	16
D	Short answer	Listening-vocabulary in syntactic context	16
E	Open response	Discourse, fluency, pronunciation, vocabulary	2

PhonePass 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, and elicit examinee responses that can be analyzed automatically to produce measures that underlie facility with English, including fluency, listening, vocabulary, accurate repetition, and pronunciation.

2.2 Item Design Specifications

All item material was crafted specifically for the test, but 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: SET-10 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). 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. The Switchboard corpus is available from the Linguistic Data Consortium (http://www.ldc.upenn.edu).

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) as, for example, "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 in speaking rate and pronunciation clarity between speakers and items is permitted.

World Knowledge and Cognitive Load: Candidates should not need specialized world knowledge or familiarity with local cultural referents to answer items correctly. SET-10 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 6 seconds. SET-10 items should not require unusual insight or feats of memory.

3. 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 processing operations have a parallel in linguistic theory, while others do not. Adapting from the model proposed by W. J. M. Levelt in his book *Speaking* [5], we can posit the psycholinguistic processing steps shown below in Figure 1. In a conversation, to understand what is said, 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 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.

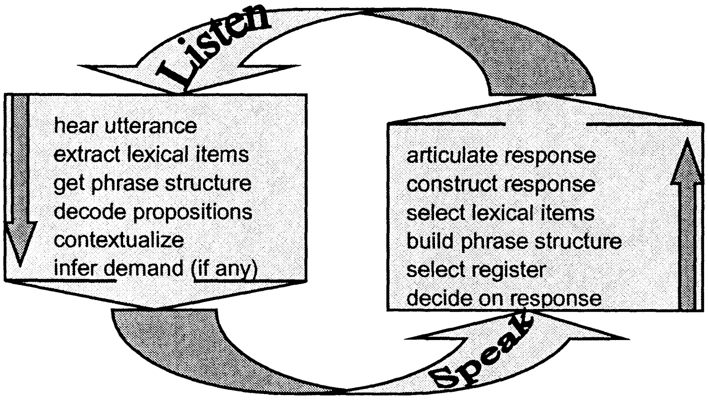

***Figure 1**: Internal processing flow in a model of speaking and listening.*

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 these basic components of psycholinguistic processing. Because the task items of the SET-10 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 the figure, except the register selection and the contextualization.

The PhonePass test is a direct test of facility with spoken English materials. Performance on the PhonePass tasks provides evidence of ability to participate in English conversation to the extent that the PhonePass tasks require many of the same skills that are used in natural conversation. Following the theoretical framework of Levelt [5], the corresponding skills include:

Generation/Speaking elements	**Sampled in test items**
1. conceive message	all items
2. select register	—
3. build phrase structure	long repeats, questions
4. select lexical items	opposites, questions
5. encode response	repeats, opposites, questions
6. articulate response	all items

Listening/understanding elements	**Sampled in test items**
1. hear utterance	all items
2. recognize lexical forms	repeats, opposites, questions
3. extract linguistic structure	long repeats, questions
4. decode propositions	questions, opposites
5. contextualize	—
6. infer demand	all items

It can be seen from this table that all of the processing steps in listening and speaking, except register selection and contextualization, are exercised in the test items. Selecting a register and contexualizing utterances are operations that can be included in a PhonePass test, but would require more elaborate items than have been included in SET-10.

4. SCORING

4.1 General Approach to Scoring

The PhonePass SET-10 Score Report presents an overall score and five component subscores. The *Overall* score for a SET-10 test represents a measure of the examinee's facility in spoken English. It is calculated as a weighted average of the five subscores, which include:

30%: **Listening Vocabulary** – understanding spoken words and producing related vocabulary
30%: **Repeat Accuracy** – repeating utterances verbatim
20%: **Pronunciation** – nativeness and intelligibility in reading aloud and repeating sentences
10%: **Reading Fluency** – rhythm, phrasing/timing in reading aloud
10%: **Repeat Fluency** – rhythmic phrasing in repeating sentences

These five subscores are relatively independent from each other, although some of them represent different measures derived from the same responses, as suggested in Table 2. For example, the Listening Vocabulary subscore is based on the responses to both the Opposites and the Short Question sections, and the Pronunciation subscore is based on responses to both the Reading and the Repeat sections of the test. Conversely, both the Repeat Accuracy and the Repeat Fluency scores are based exclusively on the responses to the Repeat items.

The subscores are of two logical types, categorical and continuous. The first type is based on the correctness of the content of the response, that is, a judgment of the number of errors based on the exact words spoken in the responses. That is, did the candidate understand the item and provide a timely and correct response. These scores (Repeat Accuracy and Listening Vocabulary) comprise 60% of the Overall score. The second type of score is based in the manner in which the response is spoken (pronunciation and fluency); these scores comprise the remaining 40% of the Overall score.

4.2 Criterion Scoring by Human Listeners

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

Ordinate asked three master graders 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, and the criteria for assigning a "no response" grade. The master graders scored a large, random sample of examinee responses, and tutored the 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, pronunciation scores for the PhonePass materials, with single-rater reliabilities between 0.82 and 0.93 for the various subskills.

4.3 Machine Scoring

All SET-10 reported scores are calculated automatically using speech recognition technology. Speech recognition in the PhonePass test is performed by an HMM based speech recognizer built using Entropic's HTK toolkit. The acoustic models, pronunciation dictionaries, and expected-response networks were developed at Ordinate using data collected during administration of PhonePass tests.

The acoustic models consist of tri-state monophone models using 5 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 PhonePass tests. The expected response networks were formed from observed responses to each item over a set of over 370 native speakers and 2700 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 section of the SET-10 test, the accuracy of the repetition can be determined giving the number of words inserted, deleted, or substituted by the candidate. These item-level scores are then combined to give a "Repeat Accuracy" component measure. This combination is done using Item ResponseTheory (IRT) such that both the difficulty of the item and the expected recognition performance of the item contributes 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 which are often misrecognized will have lower weight. Using the same method, a correct/incorrect decision for each item in Sections C and D contributes to the "Listening Vocabulary" component measure. These correct/incorrect decisions are based, in part, on observed responses to the item by native and non-native speakers.

In the Reading and Repeat sections 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 three continuous measures – Reading Fluency, Pronunciation, and Repeat Fluency. This combination is achieved using 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.

5. EVIDENCE OF VALIDITY

Many kinds of evidence can be put forward in support of an assertion that scores from a certain test are valid for a particular use. We have gathered seven kinds of evidence for the assertion that the SET-10 test provides a valid measure of facility in spoken English. This evidence includes:

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 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.

5.1 Test Samples the Performance Domain

As outlined in sections 2 and 3 above, the items of the SET-10 test 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 restricting the vocabulary 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 the limits of the performance domain. The Reading, Repeat, and Short Question items offer an opportunity for candidates to demonstrate their English skills in integrated performances that exercise most of the basic psycholinguistic components of speaking and listening.

5.2 Human Listener Skill Estimates

As introduced above in section 4.2, human listeners have judged over 129,000 individual item responses. Furthermore, human listeners have produced orthographic transcriptions of 247,000 responses during the development and validation of the SET-10 test. Applying item response theoretic analyses (Wright & Stone, 1979, [14]) to these human judgments and transcriptions, we can see from the reliability data in Table 3 that human listeners do make consistent judgments of the elemental abilities that are represented in the SET-10 subscores.

We sampled a set of 159 non-native 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 the SET-10 test. These ratings were reduced to ability subscores for each individual using a single-dimensional Item Response Theory analysis with a Rasch model. In addition, each of these individuals' responses to the vocabulary and repeat items on the test were transcribed by a human listener and the number of word errors was calculated. These results were then also analyzed using IRT to give a "human-based" ability in listening vocabulary and repeat accuracy. Finally, these five scores for each individual were combined using the same linear weighting that is used for the PhonePass Overall facility score to give a "human" Overall grade for each of the individuals.

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 that that normally found for most human rated tests. These data support the presumption that candidate responses to the items in a single, 10-minute test administration are an adequate sample of spoken material upon which to base meaningful and reliable skill estimates.

5.3 Machine Score Reliability

It is not too surprising that a machine will score a single item response consistently, but we would like to know that the PhonePass system will score many responses from a given candidate in a manner that reflects the relative consistency of those performances. The machine score reliabilities displayed in Table 2 suggest that the PhonePass system, using speech recognition technology can transcribe short utterances in SET-10 responses from non-native speakers nearly as well as a human listener can do it. Further, the data suggest that the machine's pronunciation and fluency judgments are generally similar in reliability to the level of reliability observed with a single highly trained human listener. Both the human and machine Overall scores show a reliability of 0.94. When the tests are hand transcribed and judged by a human listener and when the PhonePass system is operating without any human intervention, the reliability is equally high.

***Table 2**: Human and Machine score reliability for five subscores and the Overall score; N = 159*

Subscore	**One Human Rater**	**Machine**
Listening-Vocabulary	0.82	0.75
Repeat Accuracy	0.91	0.85
Pronunciation	0.93	0.94
Reading Fluency	0.86	0.92
Repeat Fluency	0.84	0.82
Overall	**0.94**	**0.94**

5.4 Performance of Uniform Candidate Groups

A common form of test validation is to check if 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 that have very distinct levels of the construct being measured. One would expect that the distribution of scores from a test of "proficiency in algebra" would be different for different groups; for example a group of 8 year old children, a group of high-school students, and a group of professional mathematicians.

Figure 2 displays the cumulative density distribution of SET-10 Overall scores for four groups of candidates. The score range displayed on the abscissa extends from 1 through 9, as the Overall scores are calculated inside the PhonePass system. Note that scores are only reported in the range from 2.0 to 8.0. Scores below 2.0 are reported as 2.0 and scores above 8.0 are reported as 8.0.

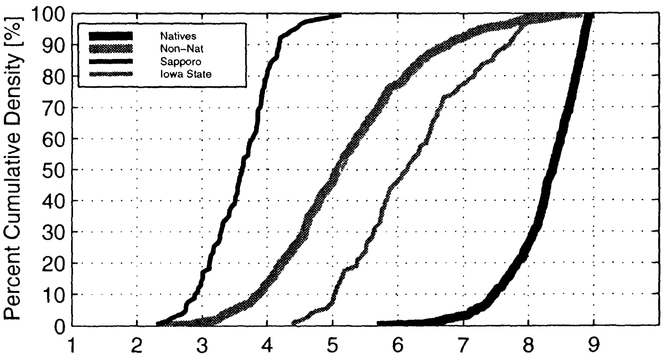

***Figure 2**: Cumulative Distribution of PhonePass Overall Score for Various Populations*

First, note the thick black line, rightmost in the figure, showing the data for a balanced set of 377 native speakers, 78% of whom got a score of 8.0. and only 2% of whom got lower than a 7.0.

The thick gray line in Figure 2 displays the cumulative score distribution for a norming group of 514 non-native speakers, ages 20 to 40, 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 of 5.1.

The two thin lines (black and gray) show two approximately homogenous populations. The thin black line, leftmost in the figure, shows a group of 96 first-year students at a Japanese university; they are all the same age, and all studied the same English curriculum for 5 years and their scores range from 2.3 to 5.0. The thin gray line is a group of 117 international graduate students, all with B.S. degrees and TOEFL scores above 550, at the point of entering an American university. Their scores range from 4.5 to 8.0.

5.5 Subscores are Reasonably Distinct

The SET-10 test 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 candidates' performance and they are identified by names that reflect the performance criteria that they are intended to measure. Thus, the Listening-Vocabulary score is derived only from items wherein the response task principally involves the immediate recognition of spoken words and ability to understand them and produce related words. The Reading Fluency score is only derived from measures of a candidate's rhythm, pace, and phrasing while reading aloud from text.

Over many candidate populations, various measures of second language ability will correlate somewhat, often with coefficients in the range 0.5 to 0.8. If subscores correlate too highly, it would be an indication that the two subscores may just be two different labels for a common ability. Note however, that 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 SET-10 subscores*

	L.-Vocab.	**Repeat Acc.**	**Read. Flu.**	**Repeat Flu.**	**Pronunc.**	**Overall**
Listening-Vocabulary	1.00	0.73	0.51	0.59	0.63	0.89
Repeat Accuracy		1.00	0.49	0.67	0.63	0.89
Reading Fluency			1.00	0.62	0.73	0.72
Repeat Fluency				1.00	0.80	0.79
Pronunciation					1.00	0.85

The machine subscores correlate with each other with coefficients in the range 0.49 — 0.80, and they correlate with the Overall score in range 0.72 — 0.89. An interesting case is displayed in Figure 3: a scatter plot of the Repeat Accuracy and Repeat Fluency scores for a norming set of 514 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.

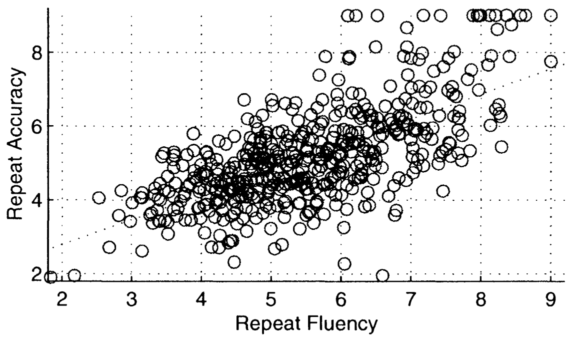

***Figure 3:** Repeat Accuracy versus Repeat Fluency; N = 514; r = 0.67*

Just as there are candidates who can read aloud but cannot speak fluently, there are candidates who can repeat a long and 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 subscore represent these differences.

5.6 Scores Correlate with Human Judgments

Scores must be reliable to be valid, but reliability alone will not establish validity. A reliable test score may just be a consistent measure of the wrong thing. Both the human scores and the machine scores for the Overall "facility" construct exhibit a reliability of 0.94, but we need to know if the machine scores actually match the human listener scores. Correlations between the machine and human subscores in Table 5 show a consistent close correspondence between human judgments and machine scores. The correlation coefficients for the two categorical scores (Listening-Vocabulary and Repeat Accuracy) are 0.89, indicating that the machine recognition algorithms that count for 60% of the score produce measures that are in close accord with the human-derived scores. The two algorithmic fluency measures match the human judgments with correlation coefficients of 0.86 and 0.87, and the automatic pronunciation score correlates with the human judgments with a correlation of 0.79.

Table 4: Correlations between machine and human scores for Overall score and subscores

Score	**Correlation**
Overall	0.94
Listening Vocabulary	0.89
Repeat Accuracy	0.89
Pronunciation	0.79
Reading Fluency	0.86
Repeat Fluency	0.87

We selected a balanced subset of 288 PhonePass testing candidates whose 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 PhonePass Overall grade. The correlation coefficient for this data is 0.94, which compares well with the single-rater reliability we observed for the human-rated Overall score of 0.94. That is, the machine grades agree with the aggregate human judgments about as well as single human raters agree with the aggregate human judgment.

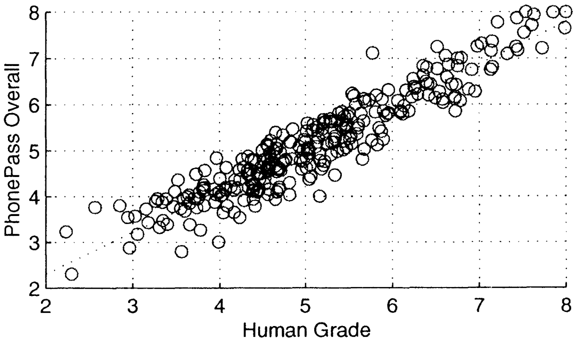

Figure 4: PhonePass Overall Facility in Spoken English vs. Human-Rater Overall Grade; N=288; r = 0.94

It is interesting to note that the close score correspondence extends over the whole range of scores. Candidates in the 2 – 4 range have difficulty producing a sentence of four words length, while candidates in the 6.5 – 8 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 PhonePass scores with other human-graded tests that have somewhat divergent target constructs.

5.7 Concurrent Validity with Tests of Related Constructs.

The predictive validity of PhonePass 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 PhonePass SET-10 tests. The correlation between the ILR OPI speaking scores and the PhonePass 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 PhonePass 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 PhonePass scoring and scores from other well-established tests of English. These results are shown in Table 5. Cascallar & Bernstein [4] conducted a study of PhonePass scoring for the US-Talk test from Regents College. The US-Talk 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. The US-Talk test shares the structure and many items with the SET-10 test, and is scored by the PhonePass system on the same facility scale.

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 PhonePass tests 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 PhonePass tests, indicate that PhonePass testing can produce scores that measure a similar underlying construct to that of TSE. Furthermore, PhonePass 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 PhonePass 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 PhonePass 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 SET-10 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 between any two tests of related but different language skills over most populations.

The study of SET-10 in relation to TOEIC was performed in Japan at the Institute for International Business Communication (IIBC). IIBC selected a stratified sample of 170 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 SET-10 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 PhonePass scoring predicts the scores of the two concurrent tests (TSE and ILR OPI) that are primarily speaking tests at or better than the level of the trained, expert raters do 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 the PhonePass SET-10 test.

6. DISCUSSION

The SET-10 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 near future, we hope to extend the range of the item types that can be automatically scored, improve the correlation between human and machine scores for pronunciation, and introduce an “intelligibility” subscore. We also hope to apply this technology to several European and Asian languages.

In the meantime, the SET-10 test is available by telephone for use from anywhere in the world, seven days a week, at any time of the day (a sample test can be found at www.ordinate.com). Test results are available within about a minute from the Ordinate Internet site. The PhonePass SET-10 scores are more reliable than best human scored tests, and the testing service may offer a convenient alternative to traditional spoken language testing methods.

7. REFERENCES

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

[2] J. Bernstein, M. Cohen, H. Murveit, D. Rtischev, & M. Weintraub, 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.

[3] J. Bernstein & D. Pisoni (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, pp. 576-579.

[4] E. Cascallar & J. Bernstein (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.

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

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

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

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

[9] G. Molholt & A. Pressler (1986) "Correlation between human and machine ratings of English reading passages", in C. Stansfield (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.

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

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

[12] L. Verhoeven (1993) "Transfer in bilingual development: The linguistic interdependence hypothesis revisited," Language Learning, vol. 44, pp. 381-415.

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

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