

## Two Experiments on Automatic Scoring of Spoken Language Proficiency

Jared Bernstein<sup>1</sup>, John De Jong<sup>2</sup>, David Pisoni<sup>3</sup>, and Brent Townshend<sup>1</sup>

<sup>1</sup>Ordinate Corporation, 1040 Noel Drive, Menlo Park, California 94025, USA

<sup>2</sup>Language Testing Services, Oranjestraat 70, 6881 SG Velp, Netherlands

<sup>3</sup>Indiana University, Bloomington, Indiana 47405, USA

jared@ordinate.com

### Abstract

New scoring methods used in the SET-10 spoken English test subsume pronunciation scoring into a more extensive analysis that estimates several skills underlying spoken language performance. The SET-10 test is described and its performance is reviewed. Two experiments explore the relation of SET-10 scoring to other conventional ways of reporting spoken language performance. Experiment 1 concerns the Council of Europe's Framework for describing second language proficiency. Samples of non-native English speech were: scored in SET-10, an automatic speaking test, and rated independently by three raters. Rater reliability in using the Council of Europe's scale and the comparability of the human and automatic measures are reported. Experiment 2 concerns the prediction of the intelligibility of non-native speech using SET-10 scoring and modifications to that scoring. A novel method for estimating intelligibility is described and preliminary results are reported. Both experiments support the notion that fully automatic methods of spoken language performance measurement can be used to predict more traditional assessments.

### Background

Since the 1960's, there have been several major (and many minor) efforts to establish curriculum-independent models of language learning and language proficiency (see North, 2000, or De Jong & Verhoeven 1992). These include the FSI and ILR scales from the U.S. government, the ACTFL Proficiency Guideline, and the Council of Europe Framework that inform human-rated testing procedures. In the same period, psycholinguists and speech engineers have developed performance measures for spoken language transmission to computers and human listeners. These measures are most often based on intelligibility or word error rate (see e.g. Miller & Isard, 1963, or Fourcin et al., 1989). A third approach to testing spoken language emerged when it was discovered that automatic signal analysis methods could measure human speaking skills (e.g. Bernstein et al., 1990, and Neumeier et al., 1996).

This paper describes two experiments that use a fully automatic test that is based on a psycholinguistic model of language processing skills, asking two questions:

1. Can such a test place candidates on a communicative-functional scale within the Council of Europe Framework?

2. Can these methods be used to predict intelligibility?

### PhonePass SET-10 Test

SET-10 stands for "Spoken English Test – 10 minutes." SET-10 is a test of speaking and listening in English that is administered over the telephone by a computer system. The test is intended for use with adult non-native speakers and the test tasks require English oral comprehension and production skills at a native-like conversational pace.

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. A fuller description of the SET-10 itself was presented at the first STiLL workshop (Townshend, et al., 1998). See also the website [www.ordinate.com](http://www.ordinate.com) for more information.

### Construct

The SET-10 construct is *Facility in Spoken English*, or ease and immediacy in understanding utterances, then formulating and producing relevant, intelligible responses at a conversational pace. SET-10 measures core linguistic skills that enable a person to participate in discussions on everyday topics among high-proficiency speakers.

Operationally, one can view the construct as defined by the procedure and scoring method. The computer system presents a series of discrete recorded prompts (out of context) to the candidate at a native conversational pace and in a range of accents and speaking styles. Each prompt requires the candidate to understand a spoken English utterance and respond to it appropriately by speaking in English. The scoring method can be summarized as follows: 60% of the score is based on the linguistic content of the candidate responses, and 40% of the score is based on the manner in which the responses are spoken, that is, pronunciation and phonological fluency.

### Machine scoring

In each part of the test, the incoming responses are recognized automatically by a speech recognition system that has been optimized for non-native speech. Recognition is performed by an HMM-based recognizer built using the HTK toolkit. Acoustic models, pronunciation dictionaries, and expected-response networks were developed in-house at Ordinate using data from PhonePass testing. The words, the pauses, the syllables, the phones, and even some subphonemic events are located in the recorded signal.

The content of the response is scored according to presence or absence of a “correct” word string for the item. This counts for 60% of the overall score, and reflects whether or not the candidate understood the prompt and responded appropriately. In this, the machine does as well or better than a naïve listener, but does not do as well as a trained listener who knows the item.

The manner scores (pronunciation and fluency) are calculated by measuring the latency of the response, the rate of speaking, the position and length of pauses, the stress and segmental forms of the words, and the pronunciation of the segments in the words. These measures are scaled according to the native and non-native distributions and then re-scaled and combined so that they optimally predict the human manner-of-speaking judgments (when the process is run on a reference set of non-native speakers). This counts for 40% of the overall score, and reflects whether or not the candidate speaks like a native (or like a favorably-judged non-native)

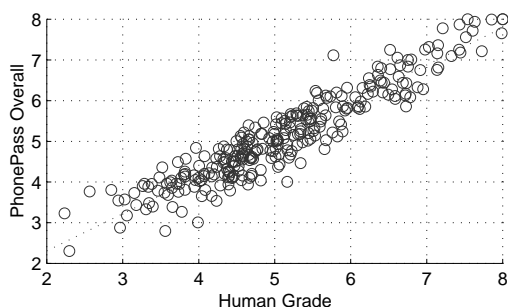
### Performance

The SET-10 Overall score has a reliability of 0.94. The correlation between SET-10 test scores and scores from three other well-known English tests is shown in Table 2. SET-10 scores correlate with scores from the ILR OPI and the TSE (human-rated oral proficiency tests) at near the level of reliability of those tests. The correlation with TOEFL scores is reasonable, given the differences in construct between the two tests.

**Table 1.** Correlation of SET-10 with Concurrent Scores

Test	Correlation	N
ILR-OPI	0.75	51
TOEFL	0.73	418
TSE	0.88	58

Figure 1 shows a scatter plot of an Overall grade produced by human expert listeners transcribing and judging responses against the SET-10 Overall grade as returned automatically by the PhonePass system. The correlation coefficient for this data is 0.94. The machine grades agree with the aggregate human judgments about as well as single human raters agree with the aggregate human judgment.



**Figure 1:** SET-10 Overall Facility in Spoken English vs. Human-Rater Overall Grade;  $N=288$ ;  $r = 0.94$

All data reported in this paper are for Version 43 scoring of the SET-10 test.

## Experiment 1

Several previous studies have reported strong correlation between PhonePass SET-10 scores and scores from speaking tests that are administered and scored by human experts. Correlations from two such concurrent validity studies are included in Table 1. Cascallar & Bernstein (2000) studied students resident in New York who took the Test of Spoken English (TSE) and the SET-10. Martinez-Scholze (1998) studied military personnel visiting Texas, who took both a government rated ILR Oral Proficiency Interview (ILR-OPI) and a SET-10 test. In both cases, the concurrent test (TSE or ILR-OPI) is designed and scored with reference to a performance framework that emphasizes communication function (not *facility*), yet the SET-10 predicts the score of the concurrent test at near the level of the concurrent test’s reliability.

Experiment 1, reported here in a preliminary form, was designed to see if the scoring of the SET-10 could be related, not to another test, but to the Council of Europe Framework itself. The goal was to establish a transformation mapping (a concordance) between SET-10 scores on the PhonePass scale and the levels designated in the Council of Europe Framework.

### Material

A scale of oral interaction proficiency according to the European Framework was constructed to incorporate elements from the Strategic, Pragmatic and Linguistic scales. The scale contains the six basic levels used in the European Framework: A1, A2, B1, B2, C1 and C2 (see the **Appendix**). Because the lowest level (A1) assumes some proficiency, a zero level was added for subjects lower than A1 and those providing no evidence of proficiency.

### Procedures

Three raters were contracted; one each from Netherlands, Switzerland and the UK. All three raters were well acquainted with the Council of Europe Framework and were retrained at the time of the experiment using two sets of examples. Raters worked independently on a practice data set, then discussed their results via e-mail and negotiated discrepancies. This was to promote similar judgment from the raters.

The response sample for each rater consisted of 131 subjects responding to the two open ended questions presented at the end of the SET-10 test. The selection of subjects was made to be substantially overlapping between raters resulting in 101 subjects being rated by all three raters and 30 subjects being rated by two. To measure intra-rater consistency, raters were presented with about 30% of the samples twice. The whole procedure was fully automatic. Raters called into the database via the telephone at their own convenience and rated as many samples per call as they wished. Raters listened to samples and entered their ratings on the telephone keypad.

Data were analyzed using the multifaceted Rasch model that allows estimating rater severity, subject ability, and item difficulty in a unified analysis. The model assumes a single underlying dimension, where values along this dimension are expressed on a logit scale (Wright & Stone, 1979).

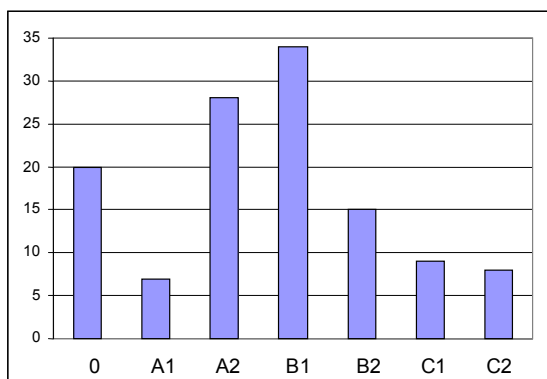
**Results**

Table 1 shows the degree of agreement among raters with respect to the placement of the cut-offs between adjacent levels of the Council of Europe Framework.

**Table 2. Cut-off estimates for three raters, in logits.**

Levels	Rater 1	Rater 2	Rater 3
C2	4.14	4.57	3.84
C1	2.83	2.35	2.65
B2	0.91	0.61	1.83
B1	-1.53	-1.15	-0.10
A2	-2.70	-2.84	-3.00
A1	-3.36	-3.59	-5.24
0	-	-	-

Table 1 also shows that raters differ in overall severity and in their estimates of the length of the scale. Rater 1 estimates the difference between the highest and the lowest cut-off as 7.5 on the logit scale, while rater 3 uses more than 9 logits. Raters are in quite good agreement on the placement of the cut-offs between levels A1 and A2 and between levels B2 and C1, achieve moderate agreement on the cut-off between levels C1 and C2, but differ significantly in their estimates for the cut-offs 0/A1, B1/B2 and B2/C1. The discrepancy, however, is never larger than the difference between two levels, which in practice means that no subject was placed further than one level apart by any two judges.

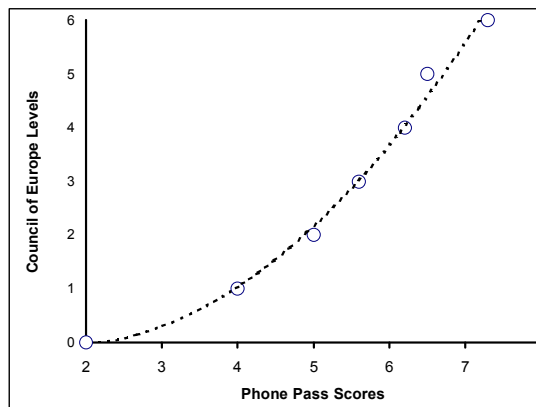


**Figure 2. Subject Distribution by Council of Europe Level**

Figure 2 presents the distribution of subjects over the Council of Europe levels using cut-offs averaged over the three raters. Figure 2 shows that a substantial number of subjects produced responses placing them at the zero level. Also assuming a normal distribution for the remainder of subjects there appears to be overrepresentation of subjects at A2, B1 and C2. Overrepresentation at level C2 corresponds to the selection of subjects with almost 10%

native speakers. Subject average level was estimated at level A2 with a standard deviation of 1.4 levels. Reliability of these estimates was 0.95.

*Relation of Council of Europe Levels to Set-10 scores*



**Figure 3. Score transformation from PhonePass Set-10 scores to Council of Europe levels**

Figure 3 shows a mapping of average subject results on the European level against the average Set-10 scores for these 131 subjects. A polynomial curve has been drawn to best fit the data from these two measures.

Figure 3 suggests that the following transformations may be warranted from the results of Experiment 1:

PhonePass scores

- from 2.0 to 3.9 predict level “below level A1,”
- from 4.0 to 4.9 predict level A1,
- from 5.0 to 5.5 predict level A2,
- from 5.6 to 6.1 predict level B1,
- from 6.2 to 6.7 predict level B2,
- from 6.8 to 7.2 predict level C1,
- from 7.3 to 8.0 predict level C2.

**Experiment 2**

People who are responsible for interpreting test results and using scores as part of a decision process want scores in a form that can be understood. Many standardized tests (e.g. TOEFL, SAT, SET-10) report performance on a numeric scale that needs to be understood in relation to the required activities that a candidate will be expected to participate in.

Tests that are framed in terms of communication functions should produce scores that are self-explanatory, which is why the map in Figure 3 may be helpful to score users: it points to a scale of descriptors (in the Appendix) that indicate how well a candidate can communicate in English by speaking. On the other hand, a numeric scale that can be directly interpreted might be more useful in some circumstances, especially if the numeric scores are anchored to a known population in an understandable way.

Experiment 2, reported here in a preliminary form, was designed to support the use of SET-10 scoring to predict how intelligible a non-native speaker will be to a particular

population of listeners – undergraduate students attending a university in the United States.

### Materials

A balanced set of 461 test-takers was assembled from a large database that archives SET-10 performances. The test-taker set was balanced for gender, and distributed over many native languages. Of the 461 test takers, 71 (15%) were native speakers of English. From these 461 test performances, 5585 response tokens (about 12 per test-taker) were selected for presentation to naïve native listeners. The response tokens were recordings of single-sentence utterances made in response to one of 246 test items selected from the first two parts of the SET-10 test (Readings or Repeats).

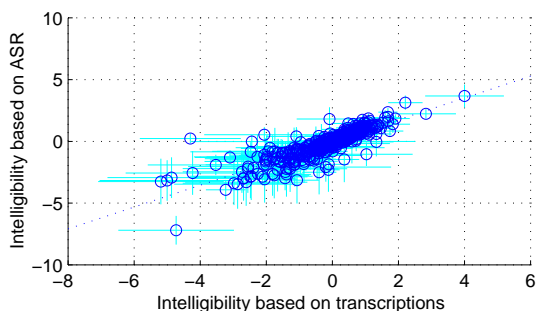
### Procedure

The naïve native-English listeners were a group of 150 undergraduate student volunteers at Indiana University. Each listener called into the PhonePass system and was presented with 200 response tokens. The listeners were instructed to listen carefully to each response and to repeat it verbatim into the phone. Listeners had no option to re-hear the recording. The 200 test-taker recordings were selected for presentation to the naïve listeners in such a way that the listeners did not hear more than one response token to a given SET-10 test item.

The naïve listeners produced a set of 29,712 usable responses. These listener responses were then used to estimate the intelligibility of the test-takers. All 29,712 listener responses were transcribed by automatic speech recognition (ASR), and a subset of 8,268 of the listener responses were also transcribed by a human operator.

### Results

Intelligibility was measured as word error rate (WER). WER was calculated for a test taker by comparing the words found in the *listener's* spoken response to the words in the test-taker's original response token. WER was implemented as the number of substitutions, deletions and insertions needed to match two word strings, with leading and trailing material ignored.

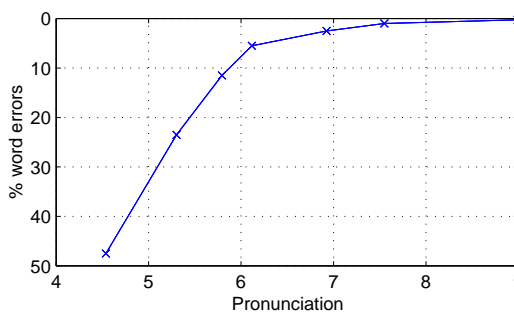


**Figure 4:** Word Error Rate, in Logits, based on Human vs. Machine Transcriptions;  $N = 443$ ;  $r = 0.86$ .

To establish that WER can be estimated accurately based on automatic recognition of the listener's responses, we compared the WER estimates for 443 test-takers for whom

we had sufficient naïve listener responses that had been transcribed by both humans and by ASR. The reliability of the WER done by ASR was 0.80 and that done by human operators was 0.78, and the correlation between the two estimates was 0.86. Figure 4 shows a scatter plot of the two estimates for the 443 test-takers.

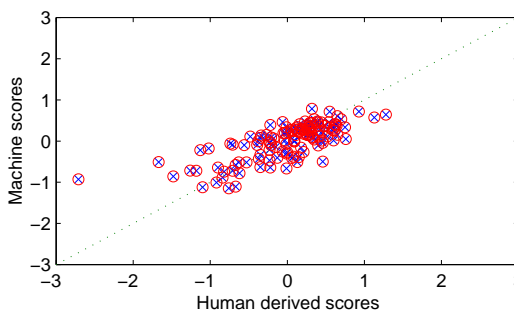
For a set of 459 test-takers, the correlation between the SET-10 Overall scores with the WER intelligibilities is 0.61. Using the SET-10 pronunciation subscore only, the correlation with WER is 0.65. If we calculate the expected pronunciation score as a function of WER, then we get the curve shown in Figure 5. Note that the average WER for self-reported native speakers was 6%.



**Figure 5:** Trend of SET-10 Pronunciation Score in relation to Word Error Rate.

The SET-10 Pronunciation subscore is based on a non-linear combination of measures of the acoustic speech signal that has been optimized to match human judgments of pronunciation quality, not intelligibility. In trying a preliminary re-combination of the base measures to predict WER, we have found, so far, that we can increase the correlation between the machine scores and the listener-derived WER scores to 0.75.

Figure 6 shows a scatter of machine predicted intelligibility scores vs. intelligibility estimates from an analysis of the listener responses. The plot is for a set-aside test set (one third of the data).



**Figure6:** Machine-predicted WER vs. Listener-based WER;  $N = 153$ ;  $r = 0.75$ .

### Discussion

The two experiments have provided preliminary answers to the questions posed at the beginning of the paper. Yes, the

automatic Overall scores from SET-10 tests can place candidates on a functional communication scale that is compatible with the Council of Europe Framework. The procedure used to score subjects responding to the two open-ended PhonePass Set-10 questions on the Council of Europe scale produced reliable estimates of the subjects' position on this scale. These estimates showed reasonable correspondence with the automatically generated scores for these subjects in responding to the remaining items within the PhonePass Set-10 spoken English test.

Initial work on predicting intelligibility (calculated as WER) has also shown positive results. There is a relatively smooth, if noisy, relation between SET-10 pronunciation scores and WER, and preliminary experiments show promising improvement in the predictive relationship.

Tech. and Language Learning (STiLL 98), Marholmen, Sweden, May 1998, pp. 179-182.

B. Wright & M. Stone (1979) *Best Test Design: Rasch Measurement*. Chicago: Mesa Press.

### References

- J. Bernstein, M. Cohen, H. Murveit, D. Rtischev, and 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.
- E. Cascallar & J. Bernstein (2000) "The Assessment of Second Language Learning as a Function of Native Language Difficulty Measured by an Automated Spoken English Test", Paper presented at the American Assoc. Applied Linguistics Annual Meeting, Vancouver, Canada, March.
- J. De Jong & L. Verhoeven (1992) Modeling and assessing language proficiency. In: L. Verhoeven and J.. de Jong (eds) *The Construct of Language Proficiency*. Amsterdam: John Benjamins
- A. Fourcin, G. Harland, W. Barry & V. Hazan (1989), *Speech Input and Output Assessment*. NY: John Wiley & Sons.
- J. Martinez-Scholze (1998) "The PhonePass Project", unpublished memo, DLI English Language Center, Lackland Airforce Base, Texas.
- G. Miller & S. Isard (1963) "Some Perceptual Consequences of Linguistic Rules," *JVLVB* (2) pp. 217-228.
- L. Neumeyer, H. Franco, M. Weintraub, and 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.
- B. North (2000) *The Development of a Common Framework Scale of language Proficiency*. New York, NY: Peter Lang.
- B. Townshend, J. Bernstein, O. Todric, and E. Warren (1998): "Estimation of Spoken Language Proficiency", Proc. ESCA Workshop on Speech

**Appendix: Council of Europe Speaking descriptors.**

<b>C2</b>	<p><b><i>Conveys finer shades of meaning precisely and naturally.</i></b></p> <p>Can express self spontaneously at length with a natural colloquial flow. Consistent grammatical and phonological control of a wide range of complex language, including appropriate use of connectors and other cohesive devices.</p>
<b>C1</b>	<p><b><i>Shows fluent, spontaneous expression in clear, well-structured speech.</i></b></p> <p>Can express self fluently and spontaneously, almost effortlessly, with a smooth flow of language. Clear, natural pronunciation. Can vary intonation and stress for emphasis. High degree of accuracy; errors are rare. Controlled use of connectors and cohesive devices.</p>
<b>B2</b>	<p><b><i>Relates information and points of view clearly and without noticeable strain.</i></b></p> <p>Can produce stretches of language with a fairly even tempo; few noticeably long pauses. Clear pronunciation and intonation. Does not make errors which cause misunderstanding. Clear, coherent, linked discourse, though there may be some "jumpiness."</p>
<b>B1</b>	<p><b><i>Relates comprehensibly main points he/she wants to make on familiar matters.</i></b></p> <p>Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair may be very evident. Pronunciation is intelligible even if a foreign accent is sometimes evident and occasional mispronunciations occur. Reasonably accurate use of main repertoire associated with more predictable situations. Can link discrete, simple elements into a connected, sequence.</p>
<b>A2</b>	<p><b><i>Relates basic information on, e.g. work, background, family, free time etc.</i></b></p> <p>Can be understood in very short utterances, even though pauses, false starts and reformulation are very evident. Pronunciation is generally clear enough to be understood despite a noticeable foreign accent. Uses some simple structures correctly, but still systematically makes basic mistakes. Can link groups of words with simple connectors such as "and," "but" and "because".</p>
<b>A1</b>	<p><b><i>Makes simple statements on personal details and very familiar topics.</i></b></p> <p>Can manage very short, isolated, mainly pre-packaged utterances. Much pausing to search for expressions, to articulate less familiar words. Pronunciation is very foreign.</p>
	<p><b><i>Candidate performs below level defined as A1</i></b></p>
	<p><b><i>Candidate's response cannot be graded:</i></b></p> <p>Insufficient evidence to decide on score category</p> <p style="text-align: center;">Silence, irrelevant or unintelligible material</p>