

Validation Summary for PhonePass™

SET-10



About This Report

This document provides a brief overview of the PhonePass Spoken English Test-10 (SET-10) for users who are responsible for interpreting results of the test. It includes: 1) an overview of the design and development of the PhonePass test; 2) the logic used in automated scoring; and 3) validity evidence for the use of PhonePass scores.

PhonePass SET-10 Overview

The PhonePass SET-10 is an automated test that measures speaking and listening skills during a 10-minute interaction over the telephone. The test is continuously available from any telephone and is scored automatically by a computer-based system. Scores are based on the exact words used in spoken responses, as well as the pace, fluency, and pronunciation of those words in phrases and sentences.

What It Measures

The PhonePass SET-10 test measures facility in spoken English, which includes the ease and immediacy in understanding and producing basic conversational English. During the course of a conversation, a person has to track what is being said, extract meaning in real time, and formulate and produce relevant, intelligible responses at a conversational pace. PhonePass SET-10 measures core skills that enable a person to understand spoken language about everyday topics and respond intelligibly at a conversational pace.

Its Uses

PhonePass testing can be an appropriate element in screening, ranking or qualifying non-native speakers of English who have at least basic reading skills. Example applications include:

- Pre-employment screening of job applicants
- Placement in conversational language classes
- Screening international students for teaching positions
- Selection of international candidates for professional development seminars

The PhonePass SET-10 test is not intended to differentiate higher ranges of language competence such as persuasiveness, discourse coherence, or facility with subtle inference or social nuance. Moreover, examinees with significant hearing impairment should not be tested using any telephone-mediated test.

How It Is Administered

PhonePass testing is administered over the telephone with a test form to which examinees may refer. The form consists of a single sheet of paper with general instructions on one side and specific instructions and examples for each part of the test on the other side. Test forms are available with the general-instructions side presented in one of several languages, while the test side of the form (with specific instructions and examples) is in English.

The test form is given to the examinee at least 5 minutes before the test begins. During this period, the examinee is given the opportunity to read both sides of the test form, to ask questions and refer to any material such as a dictionary. As the examinee interacts with the PhonePass system during test administration, an examiner voice speaks all the instructions for the various parts of the test.

These spoken instructions are also printed verbatim on the test form. Each test item requires the candidate to understand a spoken utterance and speak in response to it. Test items are presented in various voices that are distinct from the examiner voice.

Test Content

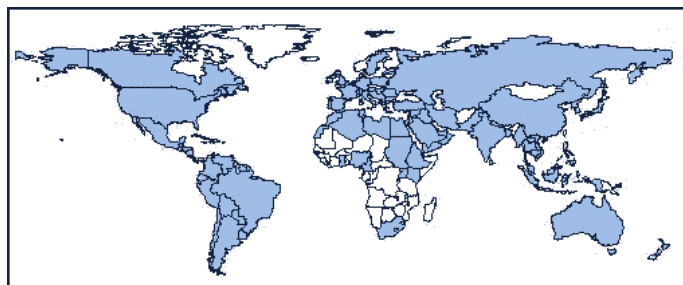
PhonePass SET-10 consists of 58 items that are presented in five separate sections (Parts A through E). Each of the five parts presents the examinee with a different task type: reading aloud, repeating sentences, saying opposite words, giving short answers to questions, and responding to open questions. In Part A, Reading, examinees are instructed to read particular sentences from among a set of numbered sentences printed on the test form. In Parts B through E, the item materials are presented by voice only.

The 58 items consist of: 8 Read-Aloud items, 16 Repeat-Sentence items, 16 Say-Opposite items, 16 Short-Answer items, and 2 Open-Question items. In scoring, there is exactly one correct word sequence expected in response to the Reading and Repeat items. Expert judgement was used to define correct answers to the Opposite and Short-Answer Question items. Most of these items have multiple answers that are accepted as correct.

Development of PhonePass SET-10

PhonePass test content covers a broad range of skill levels and skill profiles, and provides measures of fluency, listening, vocabulary, recitation, and oral reading ability in English. Lexical and stylistic patterns of actual conversation have been used in developing all item material.

Figure 1. The native languages used in the PhonePass development and validation cover most of the world. Countries addressed are shaded.



To ensure conversational content, conversations from 540 North Americans were used to guide the design of test items. Conversation samples were geographically and gender balanced and represented every major dialect of American English. An independent committee of language experts reviewed all PhonePass SET-10 items for fairness and bias-free usage. Also, to assure appropriateness for people trained to a British English standard, the items were reviewed by two British linguists to ensure conformity to colloquial usage in the United Kingdom.

All opposite and short-answer questions were pre-tested on diverse samples of native and non-native speakers. All items retained in the test were answered correctly by at least 90% of the native sample.

Scoring Method

Responses to items in Parts A through D are recorded and scored by computer. Responses to items in the last part, Open Questions, are recorded but not scored by the PhonePass system. These open responses are available for human review by authorized listeners. The PhonePass SET-10 Score Report presents an overall score and five component subscores. Each score is displayed within a confidence interval and the scores are interpreted with reference to performance descriptions. (See Figure 2.)

The Overall score for SET-10 represents a measure of the examinee's facility in spoken English. It is calculated as a weighted average of the five subscores, which include:

- Listening Vocabulary-understanding spoken words
- Repeat Accuracy-repeating utterances verbatim
- Reciting/Pronunciation-ease and nativeness in reading and repeating sentences aloud
- Reading Fluency-rhythm, phrasing/timing in reading aloud
- Repeat Fluency-rhythmic phrasing in repeating sentences

All scores are reported on a scale from 2 through 8. The Listening Vocabulary and Repeat Accuracy subscores are on a logistic scale, mapped so that the median non-native score is 5.0, and the native score at the 25th percentile is 7.5. Scores for Reciting and Fluency are mapped to criteria that are described in the score reports. For example, a Reading Fluency grade between 5.9 and 7.5 is mapped to the following criterion description: "Candidate reads with acceptable rhythm and generally appropriate phrasing; some units may be too fast or too slow. Occasional hesitation, repetition, and/or imperfect word linking may produce an uneven phrasing."

Figure 2. Example PhonePass SET-10 Score Report.

PhonePass		Score Report		Spoken English Test				
Version:	calib.24.33.37							
Test Paper Number:	66114742							
Administration date:	06 JAN 1999, 20:05 - 20:15 (PST)							
Status:	Completed							
		2	3	4	5	6	7	8
OVERALL	6.9						—○	
Listening Vocabulary	6.3						—○	
Repeat Accuracy	6.6						—○	
Reciting/Pronunciation	8.0						—○	
Reading Fluency	7.3						—○	
Repeat Fluency	7.0						—○	

Speech Processing

A special-purpose speech processing system performs the recognition, alignment, and scoring of spoken responses. This system includes an HMM-based speech recognizer using acoustic models, pronunciation dictionaries, and expected-response networks that were custom built from data collected during administrations of Phone-Pass tests to over 400 native speakers and over 3500 non-native speakers of English.

Machine Scoring

PhonePass test scores combine component measures that operate by two techniques: analysis of correct/incorrect

responses and function approximation using numerical output from the speech processing system.

The base measures are combined into three measures that are scored on a scale from 2 through 8 and include Reading Fluency, Reciting/Pronunciation, and Repeat Fluency.

Validity

Prototype versions of PhonePass SET-10 were administered in a series of validation studies to over 4000 native and non-native speakers. The native norming group (NG) comprised 377 educated adults, geographically representative of the U.S. population and from 18 to 50 in age. It had a female/male ratio of 60/40, and was 18% African-American.

The non-native norming group (NNG) consisted of 519 callers, including native speakers of 40 different languages. (See Figure 1.) The non-native norming group was selected from a larger group of more than 3500 non-native callers. For the NNG, the language distribution is similar to the TOEFL (Test of English as a Foreign Language by ETS) population, with Chinese, Spanish, Japanese, French, Korean, Italian, and Thai each represented by more than 15 speakers. Their ages range from 17 to 63, and the female/male ratio is 50/50. Many of the analyses reported below use a human-graded NNG (H-NNG) sample, which comprises the 365 NNG callers whose responses were graded by both human listeners and the PhonePass system.

Human Scoring

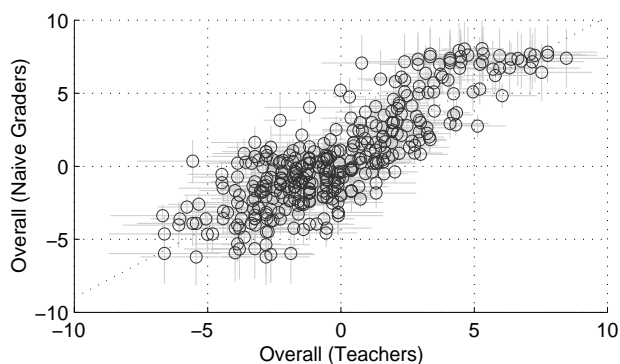
Three master graders developed, applied, and refined definitions of three scoring rubrics: fluency, pronunciation, and overall conversational skill. The master graders were expert linguists active in teaching and evaluating spoken English. The rubrics include definitions of the component skills at six levels of performance and criteria for assigning a “no response” grade. The master graders tutored other

human graders in the logic and methods used in the criterion grading.

Human graders assigned over 22,000 scores from hundreds of different examinees. Item response analysis of the human grader scores indicates that human graders produce relatively consistent grades for fluency, pronunciation, and conversational skill, with inter-grader reliabilities between 0.82 and 0.86. For the set of eight graders, the overall judgement of conversational skill based on examinee responses to open questions had a reliability of 0.93.

A second measure of human grader reliability was calculated by dividing the eight graders into two groups of four graders and correlating the grades from the two groups. One group was comprised of four professional teachers (including two of the master graders), and the other group of four 'naïve' graders, including two students and two musicians. The aggregated grades of these groups had reliabilities of 0.83 and 0.87, respectively. The inter-group score comparison is displayed in Figure 3, showing a correlation of 0.83.

Figure 3. Comparison of aggregate conversational skill grade for teachers and naïve graders. N = 375 and a correlation of $r = 0.83$.



The data presented in Figure 3 indicate that conversational skill can be judged fairly reliably by expert and naïve graders from 30-second responses to open questions that solicit an opinion. Further, the data suggest that the ordinal and interval properties of relatively naïve native judgements are generally in accord with expert judgements.

SET-10 Scoring Precision and Reliability

For a non-native norming sample (NNG) of 519 callers, the SET-10 Overall scores have a mean of 5.3 and a standard deviation of 1.23. The standard error of the overall score is 0.37.

Table 1 displays reliabilities for the component subscores as calculated from human transcriptions and judgements, and as calculated by the SET-10 scoring algorithms.

TABLE 1. Reliability Analysis for Human and PhonePass SET-10 Scoring. N = 365 (Human), N = 523 (SET-10).

Subscore	Human Score	SET-10 Score
Listening Vocabulary	.83	.80
Repeat Accuracy	.83	.85
Reciting/Pronunciation	.87	.93
Reading Fluency	.80	.93
Repeat Fluency	.83	.87
Overall	.88	.91

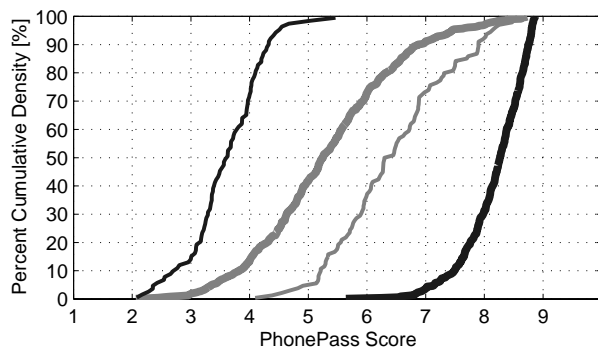
Machine-score and human score reliabilities in Table 1 are for the H-NNG group. The lower machine score reliability for Listening Vocabulary (0.80 versus 0.83) is due to the lower accuracy of the machine recognition of responses compared to human transcriber performance. The human grade reliability for Repeat Accuracy is lower than the corresponding machine grading. This is due to a slightly smaller number of responses having been transcribed by hand than were machine recognized. The higher reliability of the machine scores for Reciting and the fluency scores may be due to the uniformity of the machine's scoring performance in comparison to the human graders who sometimes conflate fluency and pronunciation in recitations and readings.

Native and Non-Native Group Performance

The performance of particular well-defined groups is a diagnostic characteristic of a test. Figure 4 presents the cumulative distribution of overall scores for four groups of examinees. The two heavy lines show the cumulative overall score distributions for the non-native norming group

(in bold light-gray) ranging from about 2 to about 9, and the native group (in bold dark-gray) ranging from about 6 to 9. The leftmost fine dark line, ranging from about 2 to 5, is a group of first year students at a Japanese college. The other fine gray line, ranging from about 4 to 9, is a group of 117 newly-arrived international graduate students at an American state university in the Midwest. All 117 of these graduate students had TOEFL scores above 500, and all but six had scores above 560.

Figure 4. Cumulative Density Functions of PhonePass SET-10 Overall scores for: the Native and Non-Native Norming Groups, Japanese undergraduates, and international graduate students entering a Midwestern state university.



Note that the range of scores displayed in Figure 4 is from 1 through 9, whereas the PhonePass system only reports scores from 2 through 8 for the SET-10 test. All scores outside the 2-8 range are deemed to have saturated the intended range of the test and are reported as 2 or 8. Only 5% of the native sample scored below 7.0, and only 10% of the non-natives scored above 7.0. The first-year English students at the Japanese university show a narrow range of scores, consistent with their relatively uniform degree of training. The international graduate students entering one university in the United States show a distribution of relatively high scores, consistent with the criteria and procedures used to select them.

SET-10 Subscore and Human Overall

Intercorrelations

Test subscores correlate with each other to some extent by virtue of presumed general covariance within the examinee population between different component elements of spoken language skill. A range of correlation values are found among the machine-generated subscores and between the machine subscore and the human subscores. These values are displayed in the following table for a group of 365 non-natives (H-NNG). Note that the columns are labeled with numbers that correspond to the numbered subscores labeled in the rows. Scores 1 through 6 are machine-generated scores; score 7 is based on human listening, as described in the following section.

TABLE 2. Scale Intercorrelations. N = 365 Non-Natives.

Score	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>
1. Listening Vocabulary	.72	.56	.45	.54	.85	.86
2. Repeat Accuracy		.59	.40	.63	.87	.86
3. Reciting/Pronunciation			.82	.92	.86	.71
4. Reading Fluency				.57	.70	.57
5. Repeat Fluency					.82	.69
6. PhonePass Overall						.93
7. Human Overall						

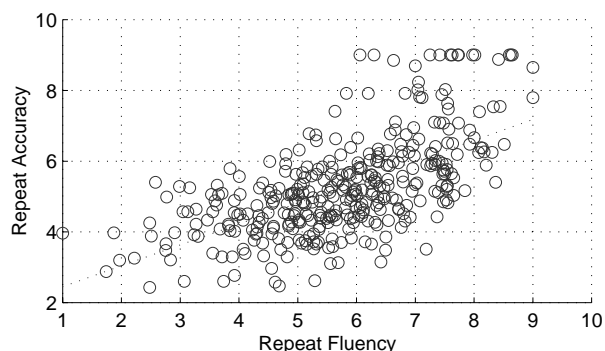
The correlation value in the lower right of Table 2 is for the human versus machine overall test scores. The value of 0.93 suggests that the overall machine scoring rates examinees in a way that is very similar to human scoring.

Table 2 shows inter-score correlation values that range from 0.40 to 0.93. The low correlation for Reading Fluency versus Repeat Accuracy ($r = 0.40$) is consistent with the fact that the scores measure distinct constructs, using different measurement methods and different sets of responses. The high correlation between Reciting and Repeat Fluency ($r = 0.92$) reflects that the two scores measure related constructs, use similar measurement methods, and are based on intersecting response sets.

Figure 5 shows the relation of two relatively independent

machine scores (Repeat Accuracy and Repeat Fluency) that are calculated from the same subset of responses. Although the two measures are derived from the same data set (Part B, *Repeats*), the two scoring algorithms extract distinct measures from the responses. The data displayed in Figure 5 is for the 365 test administrations of the non-native group, H-NNG.

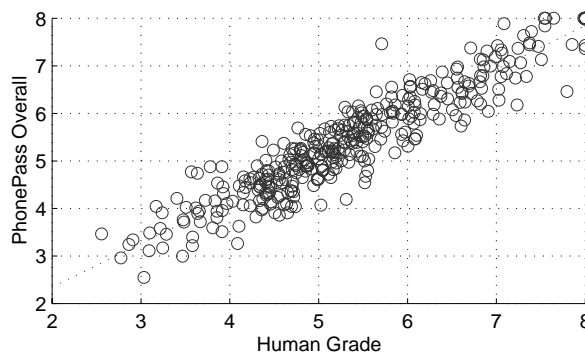
Figure 5. Machine scores of Repeat Accuracy versus Repeat Fluency for non-native callers. N = 365 and a correlation of $r = 0.63$.



Correlation Between SET-10 and Human Scores

Machine-generated SET-10 Overall scores have a correlation with human-based overall scores of $r = 0.93$. The data presented in Figure 6 show the two values for the H-NNG set of 365 examinees. The human overall grade is calculated in a manner similar to the machine Overall grade, except that human listener data is used. In the human overall grade, human transcriptions are used in place of machine-recognized transcriptions. Also, the item-level fluency and pronunciation grades were produced by listeners instead of by a machine algorithm, and the reciting grade is replaced by a human grader judgement for overall conversational skill based on listening to the open question responses.

Figure 6. PhonePass SET-10 Overall Scores versus parallel Human Scores. N = 365 and a correlation of $r = 0.93$.



Correlations with Other English Language Tests

In the validation data there were several sets of examinees who concurrently took other well-established language examinations, enabling a measure of concurrent validity of the PhonePass SET-10. Table 3 presents correlations of scores for these tests versus SET-10 Overall scores.

TABLE 3. Correlations of SET-10 Overall with Other Tests of English Proficiency.

Test	Cor- relation	Number of Examinees
TOEFL (Test of English as a Foreign Language)	.75	392
TOEIC (Test of English for International Communication)	.71	171
ILR Speaking	.77	51

The results show high correlations with these measures of related skills. A set of 392 examinees with TOEFL scores was randomly re-sampled to match the general TOEFL score distribution, establishing a correlation of 0.75 and a 75% confidence interval for the correlation of [0.71 – 0.78]. Note that the TOEFL test does not measure the same skill set as the PhonePass SET-10 test. In particular, TOEFL includes a substantial portion of items related to prescriptive and written language use. The ILR Speaking scores were taken from a diverse population of technical visitors to a U.S. government training program. Similar ILR scores (from the same site) have a reported inter-rater reliability of 0.76, which suggests a ceiling on meaningful correlation with

other measures and thus supports the inference that a correlation of 0.77 is evidence of a close relation.

Conclusions

Data from these PhonePass SET-10 studies provide evidence in support of the following conclusions:

- The test items elicit responses from which human listeners can form reliable estimates of an examinee's conversational skill components.
- The system produces precise and reliable skill estimates.
- Overall scores show effective separation between native and non-native examinees. Non-native examinees within uniform groups get generally similar scores.
- Subscores of the SET-10 are reasonably distinct and are therefore useful.
- PhonePass SET-10 scores show a high correlation with results produced by human scorers.
- PhonePass SET-10 overall scores have meaningful correlations with other tests of English proficiency.

To assure the legal defensibility of employee selection procedures, employers in the U.S. follow the Equal Employment Opportunity Commission's (EEOC's) *Uniform Guidelines for Employee Selection Procedures*. These guidelines state that employee selection procedures must be reliable and valid. The above information provides reliability and validity evidence of the PhonePass SET-10 in conformance with the prescriptions of the EEOC's *Uniform Guidelines*. Therefore, the above information provides evidence of the reliability, validity and legal defensibility of the PhonePass SET-10.

Finally, note that overall SET-10 scores have highly meaningful correlations with other measures of English language proficiency. Given the ease of administration and immediacy of test score results, the PhonePass SET-10

provides a very attractive alternative for organizations that need to assess the English language skills of individuals. This is especially the case for organizations testing on a wide-scale basis.

Further information, including sample test papers or score reports, may be obtained from the Ordinate website or at the address and phone numbers listed below.

ORDINATE

1040 Noel Drive, #102

Menlo Park, CA 94025

1-650-327-4449 phone

1-650-328-8866 fax

www.ordinate.com