SET-10®

Test Description & Validation Summary

ORDINATE

1. Introduction

The SET-10® Spoken English Test from Ordinate® Corporation evaluates the facility in oral English of people whose native language is not English. Academic institutions, international corporations and government agencies throughout the world use the SET-10 to evaluate the ability of students, staff or officers to understand spoken English and to express themselves clearly and appropriately in English. The test is intended for use with adults and with students over 15 years of age.

2. Test Description

The SET-10 is a ten-minute spoken English test for adult non-native speakers of English. The test is delivered over the telephone and scored automatically by computer. The SET-10 provides numeric scores and performance levels that describe the test taker's *facility in spoken English* – that is, the ability to understand spoken English on everyday topics and to respond appropriately at a native-like conversational pace in intelligible English.

Test Format and Administration

The SET-10 test is administered over the telephone by Ordinate's testing system. The system presents the test taker with a series of spoken prompts in English, which the test taker should respond to appropriately in spoken English. The test consists of five sections, A through E. In the first four sections, which are used to determine the test scores, the prompts and expected responses are generally single utterances. The last section contains longer questions that require more elaborate responses. In a typical test, these responses are not scored. They are used on a regular basis for validation and norming purposes.

Test administration is supported by a test paper. The test paper is a single sheet of paper with material on both sides. On the first side, the test paper displays general instructions and an introduction to the test procedures (See Appendix A.). These materials are the same for all test takers. On the second side, the test paper displays the individual test form (See Appendix B.). The individual test form is unique for each test taker. The individual test form contains the Test Identification Number, displays verbatim the spoken instructions and item examples, and presents the printed sentences that the test taker must read aloud in Part A: Reading.

It is best practice for the test administrator to give the test paper to the test taker at least five minutes before the test begins. The test taker then has the opportunity to read both sides of the test paper before the test begins and can ask questions.

After reading the test paper, the test taker calls the Ordinate testing system, is instructed by the system to enter the Test Identification Number, interacts in English with the computer over the telephone, and hangs up when finished. As the test taker interacts with the Ordinate testing system during test administration, an examiner voice speaks all the instructions for the various parts of the test.

Each test item requires the test taker to understand a spoken utterance and speak in response to it. Test items are presented in various native-speaker voices that are distinct from the examiner voice.

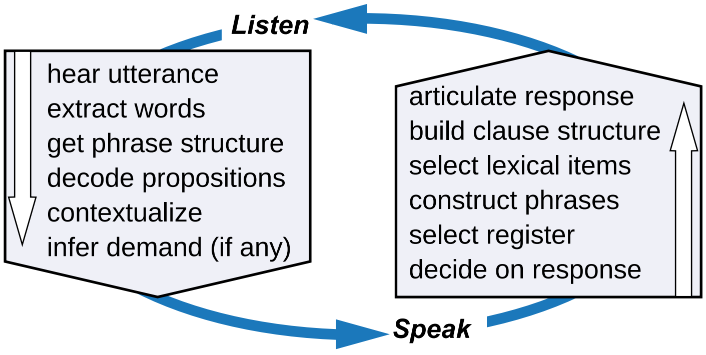
When the test administration is over, Ordinate's testing system analyzes the test taker's responses and returns an Overall score and four diagnostic subscores, usually within minutes after the call is completed. Two of the subscores reflect the content of the test taker's responses (what was said) in relation to the task presented in the item prompt. The other two subscores reflect the manner of the test taker's responses (how they were said) in terms of fluency and pronunciation quality. Test administrators and score users can view and print out test results from a password-protected section of Ordinate's website.

A typical sequence of events in the SET-10 testing process is as follows:

0. (optional, recommended) The test taker takes a practice SET-10 test.
1. The test taker gets the SET-10 test paper and reviews it.
2. The administrator answers any procedural or content questions.
3. The telephone call is made to the Ordinate testing system.
4. The test taker enters the Test Identification Number on the telephone keypad.
5. The test taker speaks in response to the content of the SET-10.
6. The telephone call is terminated, and the test taker hangs up.
7. The Ordinate system analyzes the test taker's spoken performance.
8. The Ordinate system posts the test taker's scores at www.ordinate.com.
9. The test taker or score user retrieves the SET-10 scores from www.ordinate.com.

Test Construct

The SET-10 test measures *facility in spoken English* – that is, the ability to understand spoken English on everyday topics and to respond appropriately at a native-like conversational pace in intelligible English. Another way to express the construct *facility in spoken English* is "ease and immediacy in understanding and producing appropriate conversational English." This definition relates to what occurs during the course of a spoken conversation. While keeping up with the conversational pace, a person has to track what is being said, extract meaning as speech continues, and then, on occasion, formulate and produce a relevant and intelligible response. These component processes of listening and speaking are schematized in Figure 1, adapted from Levelt (1989).



Adapted from Levelt, 1989

Figure 1. Conversational processing components in listening and speaking.

In the SET-10, the Ordinate testing system presents a series of discrete prompts to the test taker at a native conversational pace as recorded by several different native speakers, producing a range of native accents and speaking styles. These integrated "listen-then-speak" items require real-time receptive and productive processing of spoken language forms, and the items are designed to be relatively independent of social nuance and high-cognitive functions. The same facility in spoken English that enables a person to participate in everyday native-paced English conversation also enables that person to satisfactorily understand and respond to the listening/speaking tasks in the SET-10.

Explained another way, the SET-10 test measures the test taker's control of core language processing components, such as lexical access and syntactic encoding. For example, in normal everyday conversation, speakers go from building a clause structure to phonetic encoding (the last two stages in the right-hand column of Figure 1) in 40 milliseconds (Van Turennout, Hagoort, and Brown, 1998). Similarly, the other stages shown in Figure 1 have to be performed within the small period of time available to a speaker involved in everyday communication. The typical window in turn taking is about 500 milliseconds (Bull and Aylett, 1998). If language users involved in real-time communication cannot perform the whole series of mental activities presented Figure 1, both as listeners and as speakers, they will not be able to participate actively in such communication.

In this process, automaticity is required in order for the speaker/listener to be able to pay attention to what needs to be said/understood rather than to how the encoded message is to be structured/analyzed. Automaticity entails the ability to access and retrieve lexical items, to build phrases and clause structures, and to articulate these without conscious attention to the linguistic code (Cutler, 2003; Jescheniak, Hahne, and Schriefers, 2003; Levelt, 2001).

The SET-10 probes the psycholinguistic elements of spoken language performance rather than the social and rhetorical elements of communication. Because during the test this probing is performed in real time, the SET-10 measures the degree of *automaticity* in language performance. A person has to understand and produce language at some level of accuracy and fluency to participate in a spoken interchange. Since performance standards can be established for accuracy and fluency based on representative samples of language users, the SET-10 checks the level of accuracy and fluency, and at the same time directly measures the rate and level of language process control.

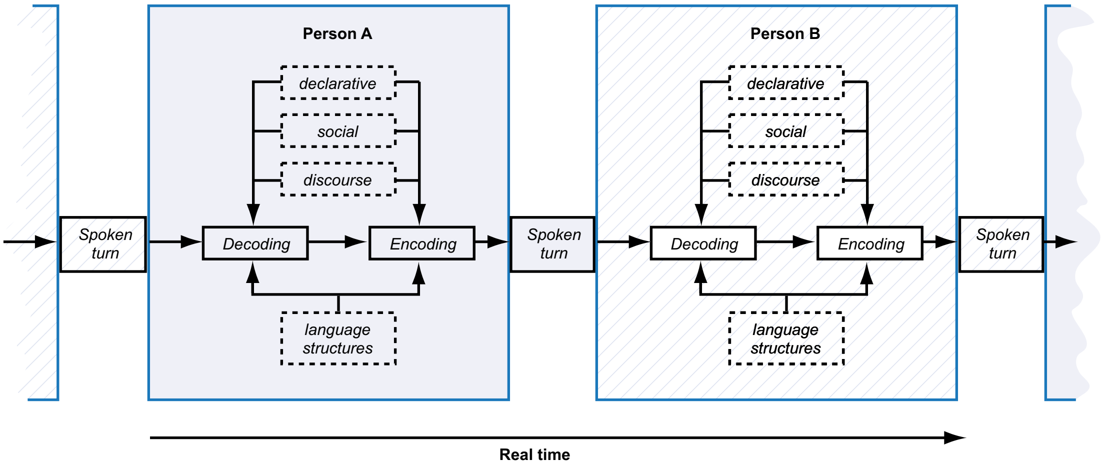


Figure 2. Message decoding and message encoding as a real-time chain-process in oral interaction.

To summarize, the SET-10 measures basic encoding and decoding of oral language as performed in integrated tasks in real time. This performance predicts a more general spoken language facility, which is essential in successful oral communication. The reason for the predictive relation between spoken language facility and oral communication skills is schematized in Figure 2. This figure puts Figure 1 into a larger context, as one might find in a socially situated dialog. The language structures that are largely shared among the members of a speech community are used to encode and decode various threads of meaning that are communicated in spoken turns. These threads of meaning that are encoded and decoded include declarative information, as well as social information and discourse markers. World knowledge and knowledge of social relations and behavior are also used in understanding the spoken turns and in formulating the content of spoken turns. However, these social-cognitive elements of communication are not represented in this model and not directly measured in the SET-10.

3. Content Design, Material and Development

Content

The SET-10 test measures both listening and speaking skills, emphasizing the test taker's facility (ease, fluency, immediacy) in responding aloud to common, everyday spoken English. All SET-10 items are designed so that both native speakers and proficient non-native speakers find them very simple to understand and to respond to appropriately. The items cover a broad range of skill levels and skill profiles. Verification of these test characteristics is described below in Section 6 on validation. The vocabulary used in the test items and responses is restricted to the most frequent words found in the Switchboard Corpus (Godfrey and Holliman, 1997), a corpus of three million words used in spontaneous telephone conversations. In general, the language structures used in the test reflect those that are common in everyday English. This includes extensive use of pronominal expressions such as “she” or “their friend” and contracted forms such as “won't” and “I'm.”

Each SET-10 item is independent of other items and presents unpredictable spoken material in English. Context-independent material is used in the test items for three reasons. First, context-independent items exercise and measure the most basic meanings of words, phrases, and clauses on which context-dependent meanings are based (Perry, 2001). Second, when language usage is relatively context-independent, task performance depends less on factors such as world knowledge and cognitive style and more on the test taker's facility with the language itself. Thus, the test performance relates most closely to language abilities and not to other test-taker characteristics that may be outside the target construct, which is facility in spoken English. Third, context-independent tasks maximize response density; that is, within a given test administration time, the test taker has more time to demonstrate performance in speaking the language. Less time is spent in developing a background cognitive schema for the task.

Both the test items presented and the expected responses are constrained to contain common English vocabulary and constructions that can be consistently understood and/or produced by at least 90% of a reference sample of educated native speakers of English. In addition, these item types maximize reliability by providing multiple, fully independent measures.

They elicit responses that can be analyzed automatically to produce measures that underlie facility with spoken English, including phonological fluency, sentence comprehension, vocabulary, and pronunciation of rhythmic and segmental units.

Test Structure

The SET-10 test consists of 61 items that are presented in five separate sections (Parts A through E). Each of the five parts presents the test taker with a different task type, as shown in Table 1.

Test Part	Task Type	Number of Items
Part A	Read-Aloud	8
Part B	Repeat-Sentence	16
Part C	Short-Answer Question	24
Part D	Sentence-Build	10
Part E	Open-Question	3

Table 1. Structure of the SET-10.

In Part A, test takers are instructed to read particular sentences from among a set of numbered sentences printed on the test paper (See Appendix B.) In parts B through E, the item materials are presented by voice only, with no direct support from the test paper. The first item response in each part of the test is not scored, and the responses to the three open questions in Part E are not scored. Thus, 54 independent responses are scored in the SET-10. In Part A, the test taker is instructed to read 8 of the 12 printed sentences. On the test paper, the 12 sentences are grouped into three related sequential groups of four in order to provide some context and limit the reasonable readings. However, the system has the test taker read the sentences in random order. The items in Parts B and D are presented in a stratified random order so that the item difficulty generally increases over the sequence of items presented. Figure 3 provides a graphical representation of the SET-10 structure, in which each vertical box represents a single item.

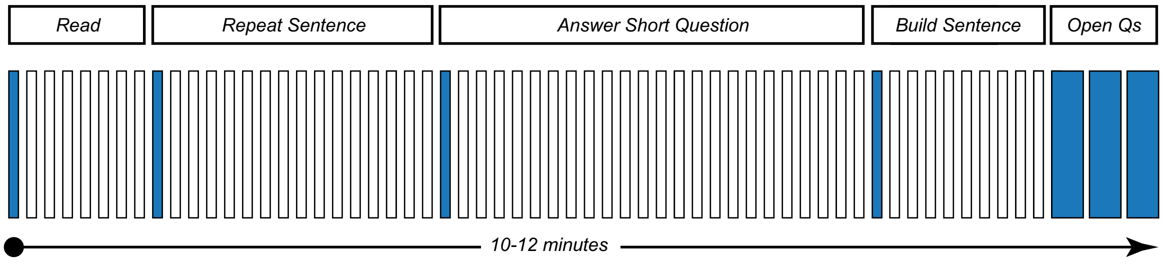


Figure 3. SET-10 task structure (blue items not scored).

SET-10 Task Design

The following subsections provide brief descriptions of the task types and the abilities required to respond to the items in each of the five parts of the SET-10 test. Item examples are also given for each task type.

Part A: Reading

In this task, test takers read printed, numbered sentences, one at a time, in the order requested. Test takers hear a request to read one of the numbered sentences printed on the test paper. Reading items are grouped into sets of four sequentially coherent sentences. The grouping helps disambiguate how each component sentence should be read.

Examples:

1. Traffic is a huge problem in Southern California.
2. The endless city has no coherent mass transit system.
3. Sharing rides was going to be the solution to rush-hour traffic.
4. Most people still want to drive their own cars, though.

The sentences are relatively simple in structure and vocabulary, and they can be read easily and in a fluent manner by literate native speakers of English. For examinees with little facility in spoken English but with some reading skills, this task provides samples of their pronunciation and reading fluency. The readings start the test because, for many test takers, reading aloud presents a familiar task, and thus this task provides a comfortable introduction to the interactive mode of the test as a whole.

Part B: Repeat

In this task, test takers repeat sentences verbatim. The sentences are presented to the test taker in order of increasing difficulty. To repeat a sentence longer than about seven syllables, the test taker has to recognize the words as produced in a continuous stream of speech (Miller & Isard, 1963). As the sentences increase in length and complexity, the task becomes increasingly difficult for speakers who are not familiar with spoken English. Highly proficient speakers of English can generally repeat sentences that contain many more than seven syllables because these speakers are very familiar with English words and phrase structures and with the common syntactic forms of English clauses in typical sentences. If a person habitually processes five-word phrases (e.g. “her really big apple tree”) as a unit, then that person can usually repeat utterances of 15 or 20 words in length. Generally, repetition of material is constrained by the size of the linguistic unit that a person can process in an automatic or nearly automatic fashion.

Examples:

War broke out.
It’s supposed to rain tomorrow, isn’t it?
There are three basic ways in which a story might be told to someone.

Part C: Short-answer questions

In this task, test takers listen to a spoken question and then answer the question with a single word or a short phrase. The questions generally present three or four (sometimes more) lexical items spoken in a continuous phonological form and framed in an English sentence structure. To respond to the question prompt, the test taker needs to identify the words in phonological and syntactic context, and infer the demand proposition. Each question asks for basic information, or for simple inferences based on time, sequence, number, lexical content, or logic. The questions do not presume any particular familiarity with specific facts of Anglo-American culture, geography, history, or other subject matter; they are intended to be within the realm of familiarity of both a typical 12-year-old native speaker of English and an adult who has never lived in an English-speaking country. Expert judgment was used initially to define correct answers to these items. Many of the items have multiple answers that are accepted as correct. All questions are pre-tested on diverse samples of native and non-native speakers. A minimum criterion for short answer items to be retained in the test is a 90% correct response rate from the native-speaker sample.

Examples:

What season comes before spring?

What is frozen water called?

Does a tree usually have fewer trunks or branches?

Part D: Sentence Builds

In this task, test takers are presented with a sequence of three short phrasal word groups. The phrases are presented in a random sequence and test takers are asked to rearrange them into a sentence. This task initially requires receptive lexical and local syntactic processing; the test taker also has to understand the possible meanings of the phrases and know how they might be likely to combine with other phrasal material. The length and complexity of the sentence that can be built is constrained by the size of the linguistic unit (e.g., one word or a three-word phrase) that a person can hold in working memory for processing.

Examples:

in / bed / stay

ralph / this photograph / could convince

we wondered / would fit in here / whether the new piano

Part E: Open Questions

In this task, test takers listen to a spoken question and then give their opinion. The questions deal either with family life or with the test taker's preferences and choices. This task is used to collect a spontaneous speech sample from the test taker. The test taker's responses are not scored at present, but these responses are available for human review by authorized listeners.

Examples:

Do you think television has had a positive or negative effect on family life? Please explain.

Do you like playing more in individiual or in team sports? Please explain.

Content Development

The SET-10 test content covers a broad range of skill levels and skill profiles, and provides measures of fluency, vocabulary, pronunciation, and sentence mastery in English. Lexical and stylistic patterns of actual conversation have been used in developing all item material. To ensure conversational content, conversations from 540 North Americans were used to guide the design of test items. Conversation samples were balanced by geography and gender and represented every major dialect of American English. In addition, steps were taken to assure that items would be appropriate for test takers trained to standards other than U.S. English; British and Australian linguists reviewed all items to ensure conformity to colloquial usage in the United Kingdom and Australia.

During the development of earlier versions of the SET-10, the test was administered in a series of validation studies to over 4,000 native and non-native speakers. Non-native speakers were sampled from a range of countries in Europe and Asia (Ordinate, 2000). The latest version of the SET-10 test builds on previous versions of the test. As part of the development of the latest version of the test, response data were collected from a sample of about 775 native speakers of English, including both speakers of American English and speakers of British English and from a sample of about 600 non-native speakers of English with a variety of language backgrounds. Results from these pretests are presented in Section 6.

4. Scoring Logic

The SET-10 score report comprises an Overall score and four diagnostic subscores: Sentence Mastery, Vocabulary, Fluency[1], and Pronunciation. These scores are defined as follows:

Overall: The Overall score of the test represents the ability to understand spoken English and speak it intelligibly at a native conversational pace on everyday topics. Scores are based on a weighted combination of the four diagnostic subscores. Scores are reported in the range from 20 to 80.

Sentence Mastery: Sentence Mastery reflects the ability to understand, recall, and produce English phrases and clauses in complete sentences. Performance depends on accurate syntactic processing and appropriate usage of words, phrases, and clauses in meaningful sentence structures.

Vocabulary: Vocabulary reflects the ability to understand common everyday words spoken in sentence context and to produce such words as needed. Performance depends on familiarity with the form and meaning of everyday words and their use in connected speech.

[1] *Within the context of language acquisition, the term "fluency" is sometimes used in the broader sense of general language mastery. In the narrower sense used in SET-10 score reporting, "fluency" is taken as a component of oral proficiency that describes certain characteristics of the observable performance. Following this usage, Lennon (1990) identifies fluency as "an impression on the listener's part that the psycholinguistic processes of speech planning and speech production are functioning easily and efficiently" (p. 391). In Lennon's view, surface fluency is an indication of a fluent process of encoding. The SET-10 fluency subscore is based on measurements of surface features such as the response latency, speaking rate, and continuity in speech flow, but as a constituent of the Overall score it is also an indication of the ease of the underlying encoding process.*

Fluency: Fluency reflects the rhythm, phrasing and timing evident in constructing, reading, and repeating sentences.

Pronunciation: Pronunciation reflects the ability to produce consonants, vowels, and stress in a native-like manner in sentence context. Performance depends on knowledge of the phonological structure of everyday words.

Figure 4 shows the mapping of these subscores into the responses for the five sections of the SET-10. Each vertical rectangle represents a response utterance from a test taker. At the beginning of sections B, C, and D of the test, the test taker hears one or more examples of the task type followed by an acceptable non-native response. The three open questions at the end of the test are not scored, nor are the test taker's responses to the first item in each of the scored sections of the test. These unscored responses are shown in blue in Figure 4.

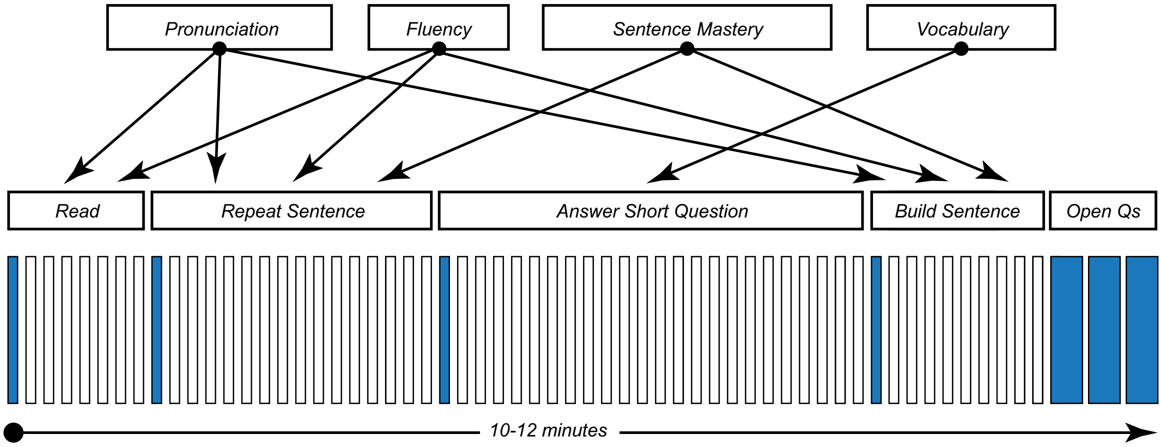


Figure 4. Relation of subscores to item types.

Among the four subscores, two basic types of scores are distinguished: scores relating to the *content* of what a test taker says (Sentence Mastery and Vocabulary) and scores relating to the *manner* (quality) of the response production (Fluency and Pronunciation). This distinction corresponds roughly to Carroll's (1961) distinction between language performance in relation to a knowledge aspect and a control aspect. In later publications, Carroll (1986) identified the control aspect as automatization, which suggests that people speaking fluently without realizing they are using their knowledge about a language have attained the level of automatic processing as described by Schneider & Shiffrin (1977).

In each section of the SET-10, each incoming response is recognized automatically by a speech recognizer that has been optimized for non-native speech. 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 the presence or absence of expected correct words in correct sequences. This content accuracy dimension counts for 50% of the Overall score, and reflects whether or not the test taker understood the prompt and responded with appropriate content. In this aspect, the machine generally does as well or better than a naïve listener, but does not generally do as well as a trained listener who knows the item.

The manner-of-speaking scores (fluency and pronunciation, or the control dimension) are calculated by measuring the latency of the response, the rate of speaking, the position and length of pauses, the stress and segmental forms of the words, and the pronunciation of the segments in the words within their lexical and phrasal context. These measures are scaled according to the native and non-native distributions and then re-scaled and combined so that they optimally predict the human judgments on manner-of-speaking (when the process is run on a reference set of non-native speakers). The manner-of-speaking scores count for the remaining 50% of the Overall score, and reflect whether or not the test taker speaks like a native (or like a favorably-judged non-native).

Producing accurate lexical and structural content is important, but excessive attention to accuracy can lead to disfluent speech production and can also hinder oral communication; on the other hand, inappropriate word usage and misunderstood syntactic structures can also hinder communication. In the SET-10 scoring logic, content and manner (i.e., accuracy and control) are weighted equally because successful communication depends on both.

5. Score Use

Ordinate endorses the use of SET-10 scores by educational and government institutions and by commercial and business organizations for making valid decisions about oral English interaction skills of individuals, provided score users have reliable evidence confirming the identity of the individuals at the time of test administration. Score users may obtain such evidence either by administering the SET-10 themselves or by having them administered by trusted third parties. In several countries, educational and commercial institutions provide such services. (Consult the International Contacts page on the Ordinate website at: http://www.ordinate.com.)

SET-10 scores can be used to evaluate the level of spoken English skills of individuals entering into, progressing through, and exiting English language courses. Scores may also be used effectively in evaluating whether an individual’s level of spoken English is sufficient to perform certain tasks or functions requiring mastery of spoken English.

The SET-10 score scale covers a wide range of abilities in using English in spoken communication. Score users must decide what SET-10 score can be regarded as a minimum requirement in their context. To establish these minimum requirements, score users may wish to consult the Ordinate SET-10 Can-Do Guide (2003), which is available from Ordinate’s Sales Department (sales@ordinate.com).

Alternatively, score users may wish to base their selection of an appropriate criterion score on their own localized research. Ordinate can provide a Benchmarking Kit and further assistance to score users in establishing criterion scores.

6. Validation

Prototype versions of the SET-10 were administered in a series of validation studies to over 4,000 native and non-native speakers. The native norming group comprised 376 literate adults, geographically representative of the U.S. population aged 18 to 50. It had a female/male ratio of 60/40, and was 18% African-American. The non-native norming group was a stratified random sample of 514 callers sampled from a larger group of more than 3,500 non-native callers. Stratification was aimed at obtaining an even representation for gender and for native language.

Over 40 different languages were represented in the non-native norming group, including Arabic, Chinese, Spanish, Japanese, French, Korean, Italian, and Thai. Ages ranged from 17 to 79, and the female/male ratio was 50/50. More information about these previous validation studies can be found in Validation Summary for PhonePass SET-10, available from Ordinate's website.

Because of the introduction of several modifications to the SET-10 in the current version a number of additional validation studies were performed. These studies used a native norming group of 775 native speakers of English, from the U.S. and the U.K. and a non-native norming group of 603 speakers from a number of countries in Asia, Europe and South America. The native norming group consisted of approximately 33% speakers from the U.K. and 66% speakers from the USA and had a female/male ratio of 55/45. Ages ranged from 18 to 75. The non-native norming group had a female/male ratio of 62/38. Ages ranged from 12 to 56.

The correlation between the current version of the SET-10 and the version for which previous validation studies were conducted is 0.98 (n=200). This suggests that many of the inferences from validation studies conducted with the previous release of the SET-10 remain warranted for the new version.

Native and Non-native Group Performance

Figure 5 presents the main results for the two norming groups. The figure shows the cumulative distribution of Overall scores for the native and non-native speakers. Note that the range of scores displayed in this figure is from 10 through 90, whereas the SET-10 scores are reported on a scale from 20 to 80. Scores outside the 20 to 80 range are deemed to have saturated the intended measurement range of the test and are reported as 20 or 80.

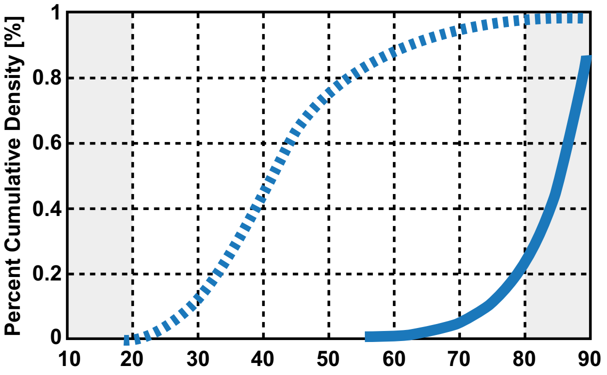


Figure 5. Cumulative density functions of SET-10 Overall scores for the native and non-native norming groups (native n=775 and non-native n=603).

The results show that native speakers of English consistently obtain high scores on the SET-10. Fewer than 5% of the native sample scored below 68. Learners of English as a second or foreign language, on the other hand, are distributed over a wide range of SET-10 scores. Note also that only 5% of the non-natives scored above 68. In sum, the Overall scores show effective separation between native and non-native test-takers.

Correlations Among Subscores

Table 2 presents the correlations among the SET-10 subscores and the Overall score for the non-native sample.

	Vocabulary	Pronunciation	Fluency	SET-10 Overall
Sentence Mastery	0.73	0.71	0.67	**0.88**
Vocabulary		0.65	0.61	**0.84**
Pronunciation			0.92	**0.92**
Fluency				**0.90**

Table 2. Correlations among SET-10 subscores for the non-native sample (n=603).

Test subscores correlate with each other to some extent by virtue of presumed general covariance within the test-taker population between different component elements of spoken language skills. The correlations between the subscores are, however, significantly below unity, which indicates that the different scores measure different aspects of the test construct, using different measurement methods, and different sets of responses.

Figure 6 illustrates the relationship between two relatively independent machine scores (Sentence Mastery and Fluency). These machine scores are calculated from a subset of responses that are mostly overlapping (Repeats and Sentence Builds for Sentence Mastery and Repeats, Sentence Builds and Readings for Fluency). Although these measures are derived from a data set that contains mostly the same responses, the subscores clearly extract distinct measures from these responses. For example, many test takers with Fluency scores in the 50-70 range have a Sentence Mastery score in the 20-40 range.

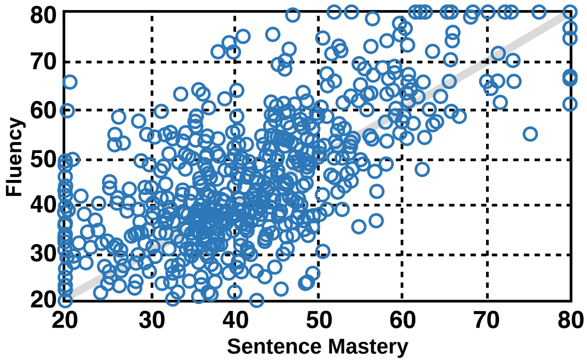


Figure 6. Machine scores of Sentence Mastery versus Fluency for the non-native norming group (n=603 and r=0.67).

SET-10 Scoring Precision and Reliability

For the non-native sample (n=603), the SET-10 Overall scores have a mean of 43 and a standard deviation of 13. The standard error of the Overall score is 2.9.

Table 3 displays reliabilities for a subset of 50 calls for which both machine scores and human scores were computed. The human scores were calculated from human transcriptions (for the Sentence Mastery and Vocabulary subscores) and human judgments (for the Pronunciation and Fluency subscores). That is, Table 3 compares the same human performances, scored by close human rating in one case and by independent automatic machine scoring in the SET-10 case. The values in Table 3 suggest that there is sufficient information in a SET-10 item response set to extract reliable information, and that the effect on reliability of scoring with Ordinate's speech recognition technology, as opposed to a careful human rating, is quite small.

Types of Score	Human Score	SET-10 Score
Overall	0.98	0.97
Sentence Mastery	0.96	0.93
Vocabulary	0.85	0.88
Fluency	0.98	0.95
Pronunciation	0.98	0.97

Table 3. Reliability analysis for human scoring (one rater) and SET-10 machine scoring (n=50).

Correlations Between SET-10 and Human Scores

Table 4 presents correlations between machine-generated scores and human scores for the same subset of 50 test-takers. The correlations presented in Table 4 suggest that the SET-10 machine-generated scores are not only reliable, but that they generally correspond as they should with human ratings. Among the subscores, the human-machine relation is closer for the content accuracy scores than for the manner-of-speaking scores, but the relation is close for all four subscores. At the Overall score level, SET-10 machine-generated scores are virtually indistinguishable from a scoring that is done by careful human transcriptions and repeated independent human judgements.

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

Table 4. Correlations between SET-10 and human scores (n=50).

The data presented in Figure 7 show human and machine scores for this subset.

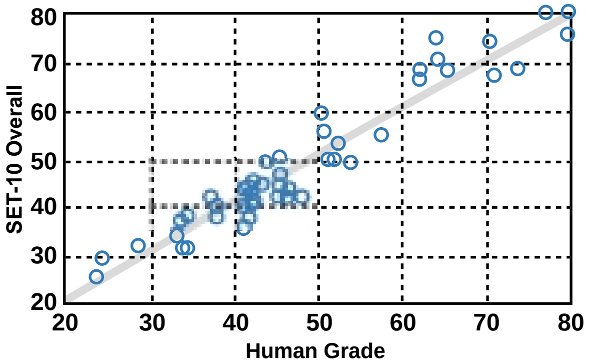


Figure 7. SET-10 scores versus human scores (n=50).

Correlations With Other English Language Tests

Over the years Ordinate and third parties have collected data on parallel administrations of the SET-10 and other well-established language examinations, enabling a measure of concurrent validity of the SET-10.

Instrument	r	n
TOEFL	0.75	392
TOEFL Reading[1]	0.64	321
TOEIC listening	0.71	171
TOEFL listening[1]	0.79	321
New TOEFL Listening[1]	0.78	321
TSE	0.88	58
New TOEFL Speaking[1]	0.84	321
Common European Framework, 1st experiment	0.84	121
Common European Framework, 2nd experiment	0.94	150
Common European Framework, 3rd experiment	0.88	303
ILR speaking	0.75	51

[1] Source: Enright, Bridgeman, and F. Cline, 2002; all other data: Ordinate Corporation

Table 5. Correlations of SET-10 with other measures.

Table 5 presents correlations of scores for these instruments with SET-10 Overall scores. The table is divided into three sections: the upper section shows data on tests of written language, which are expected to have only a moderate correlation with a speaking test such as the SET-10. The middle section shows tests of listening comprehension, which, being in the oral mode, are expected to have a somewhat higher correlation with the SET-10. The bottom section shows correlations with instruments for assessing oral skills, which focus mainly or entirely on speaking. These instruments are expected to show the highest correlation with the SET-10. The data suggest that the SET-10 measure overlaps substantially with that of other instruments designed to assess spoken language skills.

Table 5 includes data from three independent experiments conducted by Ordinate to relate the SET-10 reporting scale to an oral interaction scale based on the Common European Framework (Council of Europe, 2001). The first experiment was reported by Bernstein et al. (2000); the second experiment is reported in Ordinate (2003); and the third experiment was conducted especially for the validation of the current version of the SET-10. Responses to Open Questions from a subsample of both norming groups were assigned randomly to 6 raters who together produced 7,266 independent ratings in an overlapping design. The ratings from the two raters with the largest amount of overlapping data related to 397 responses. These raters showed perfect agreement in assigning a Common European Framework (CEF) level to 63% of the cases and differed by only one level in a further 30% of the cases. Rater agreement overall was 0.89.

Figure 8 shows the relationship between the SET-10 score and the CEF levels that became apparent from this experiment. The correlation was 0.88. The graph also shows how both instruments (the SET-10 and the CEF) clearly separate the native and non-native norming groups.

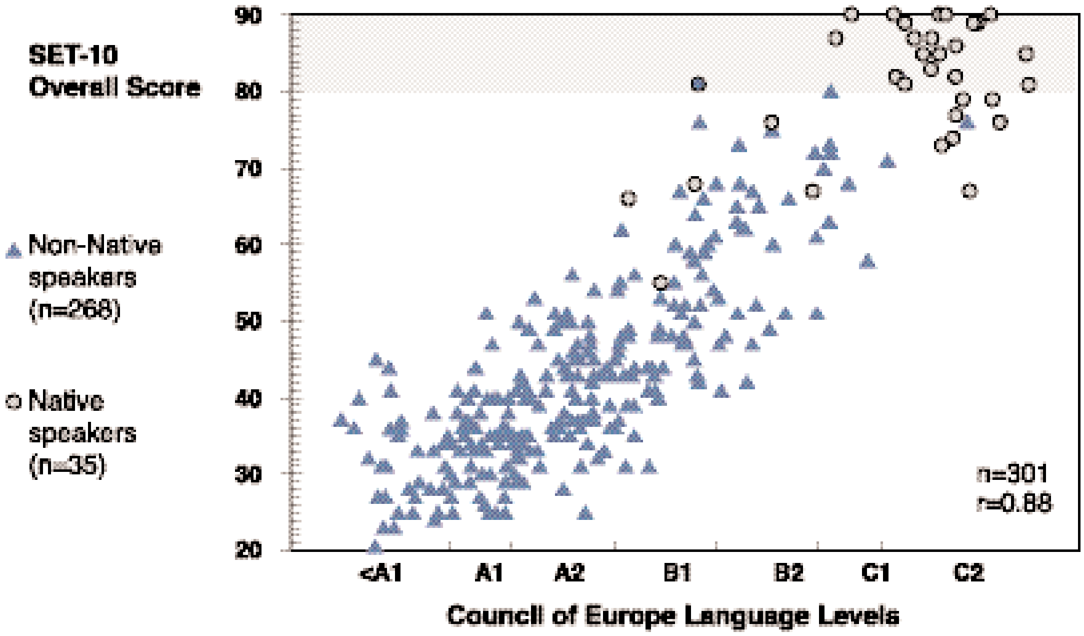


Figure 8. Correlation between SET-10 Overall score and CEF-levels.

Conclusions

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

- The Ordinate system produces precise and reliable skill estimates.
- Overall scores show effective separation between native and non-native examinees.
- Subscores of SET-10 are reasonably distinct and therefore offer useful diagnostics.
- SET-10 scores show a high correlation with human-produced ratings.
- SET-10 Overall scores have meaningful correlations with related tests of English proficiency.

To assure the 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 evidence of the reliability, validity and legal defensibility of the SET-10 in conformance with the prescriptions of the EEOC's Uniform Guidelines. Finally, note that Overall SET-10 scores have highly meaningful correlations with other measures of English language proficiency.

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

7. Ordinate Corporation: Testing and Technology

Ordinate's automated testing system was developed to apply advanced speech recognition techniques and data collection via the telephone to the evaluation of language skills. The system includes automatic telephone reply procedures, dedicated speech recognizers, speech analyzers, databanks for digital storage of speech samples, and links up to test- and language-specific procedures for scoring incoming speech data via scoring report generators linked to the Internet. The SET-10 is the result of years of research in speech recognition, statistical modeling, linguistics, and testing theory. Ordinate's patented technologies are applied to its own language tests such as the SET-10 series and also to customized tests. Sample projects include assessment of spoken languages other than English, children's reading assessment, adult literacy assessment, and collection and human rating of spoken language samples.

For Ordinate's language tests, the computer system tracks linguistic, indexical and paralinguistic characteristics of the spoken input of users, measures several speech quality aspects such as the response latency, speaking rate and pronunciation of the user, and combines that information with the relative accuracy of the linguistic content of the user's utterance to derive a set of diagnostic scores that can be combined into an overall oral interaction score.

Ordinate Corporation. Ordinate Corporation, Menlo Park, California, was founded in 1996 to develop language testing systems based on new techniques applying speech recognition to language assessment. Ordinate is the first company to develop a completely automated method for testing spoken language.

Ordinate's Advisory Board. Ordinate's Advisory Board has been set up to guide Ordinate's scientific policy for design, development, and validation of language tests. The Board includes experts from the USA and abroad in applied linguistics, speech recognition, English as a second language, psychological measurement, testing technology, and social policy. Its primary role is to help define and prioritize Ordinate's research direction. In addition, the Board critically evaluates language testing instruments and procedures developed by Ordinate and reviews development projects on an ongoing basis.

Ordinate's Policy. Ordinate Corporation is committed to the best practices in the development, use, and administration of language tests. Each Ordinate employee strives to achieve the highest standards in test publishing and test practice. As applicable, Ordinate follows the guidelines propounded in the Standards for Educational and Psychological Testing[2], and the Code of Professional Responsibilities in Educational Measurement[3]. Each employee is given a copy of the Code of Professional Responsibilities in Educational Measurement. A copy of the Standards for Educational and Psychological Testing is available to every employee for reference.

Research. In close cooperation with international experts, Ordinate conducts ongoing research aimed at gathering substantial evidence for the validity, reliability, and practicality of its current products and at investigating new applications for Ordinate technology. Research results are published in international journals and made available through the Ordinate website (http://www.ordinate.com).

8. References

Bernstein, J., De Jong, J.H.A.L., Pisoni, D., & Townshend, B. (2000). Two Experiments on Automatic Scoring of Spoken Language Proficiency. In: P. Delcloque (Ed.), *Proceedings of InSTIL2000: Integrating Speech Technology in Learning* (pp. 57-61). University of Abertay Dundee, Scotland, August, 2000.

Bull, M., & Aylett, M. (1998). An analysis of the timing of turn-taking in a corpus of goal-oriented dialogue. In: R. H., Mannell & J. Robert-Ribes (Eds.), *Proceedings of the 5th International Conference on Spoken Language Processing.* Canberra: Australian Speech Science and Technology Association.

Carroll, J.B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. *Testing.* Washington, DC: Center for Applied Linguistics.

Carroll, J.B. (1986). Second Language. In: R.F. Dillon, & R.J. Sternberg (Eds.), *Cognition and Instruction.* Orlando, FL: Academic Press.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

Cutler, A. (2003). Lexical access. In L. Nadel (Ed.), *Encyclopedia of Cognitive Science. Vol. 2, Epilepsy - Mental imagery, philosophical issues about* (pp. 858-864). London: Nature Publishing Group.

[2] The Standards for Educational and Psychological Testing (1985; 1999). Developed jointly by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). These standards are published in book format and can be ordered through the Internet link: http://www.apa.org/science/standards.html.

[3] Code of Professional Responsibilities in Educational Measurement (1995). Prepared by the NCME Ad Hoc Committee on the Development of a Code of Ethics. The full text can be found at: http://www.natd.org/Code_of_Professional_Responsibilities.html.

Enright, M.K., Bridgeman, B., & Cline, F. (2002, April). *Prototyping a Test Design for a New TOEFL.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Godfrey, J.J., & Holliman, E. (1997). *Switchboard-1 Release 2.* LDC Catalog No.: LDC97S62. http://www.ldc.upenn.edu

Jescheniak, J.D., Hahne, A., & Schriefers, H.J. (2003). Information flow in the mental lexicon during speech planning: evidence from event-related brain potentials. *Cognitive brain research, 15* (3), 261-276.

Levelt, W. J. M. (1989). *Speaking: From Intention to Articulation.* Cambridge, MA: MIT Press.

Levelt, W. J. M. (2001). Spoken word production: A theory of lexical access. *PNAS, 98* (23), 13464-13471.

Miller, G.A., & Isard, S. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior, 2,* 217-228.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning, 40,* 387-412.

Ordinate (2000). *Validation summary for PhonePass SET-10: Spoken English Test-10, System Revision 43.* Menlo Park, CA: Author.

Ordinate (2003). *Ordinate SET-10 Can-Do Guide.* Menlo Park, CA: Author.

Perry, J. (2001). *Reference and Reflexivity.* Stanford, CA: CSLI. Publications.

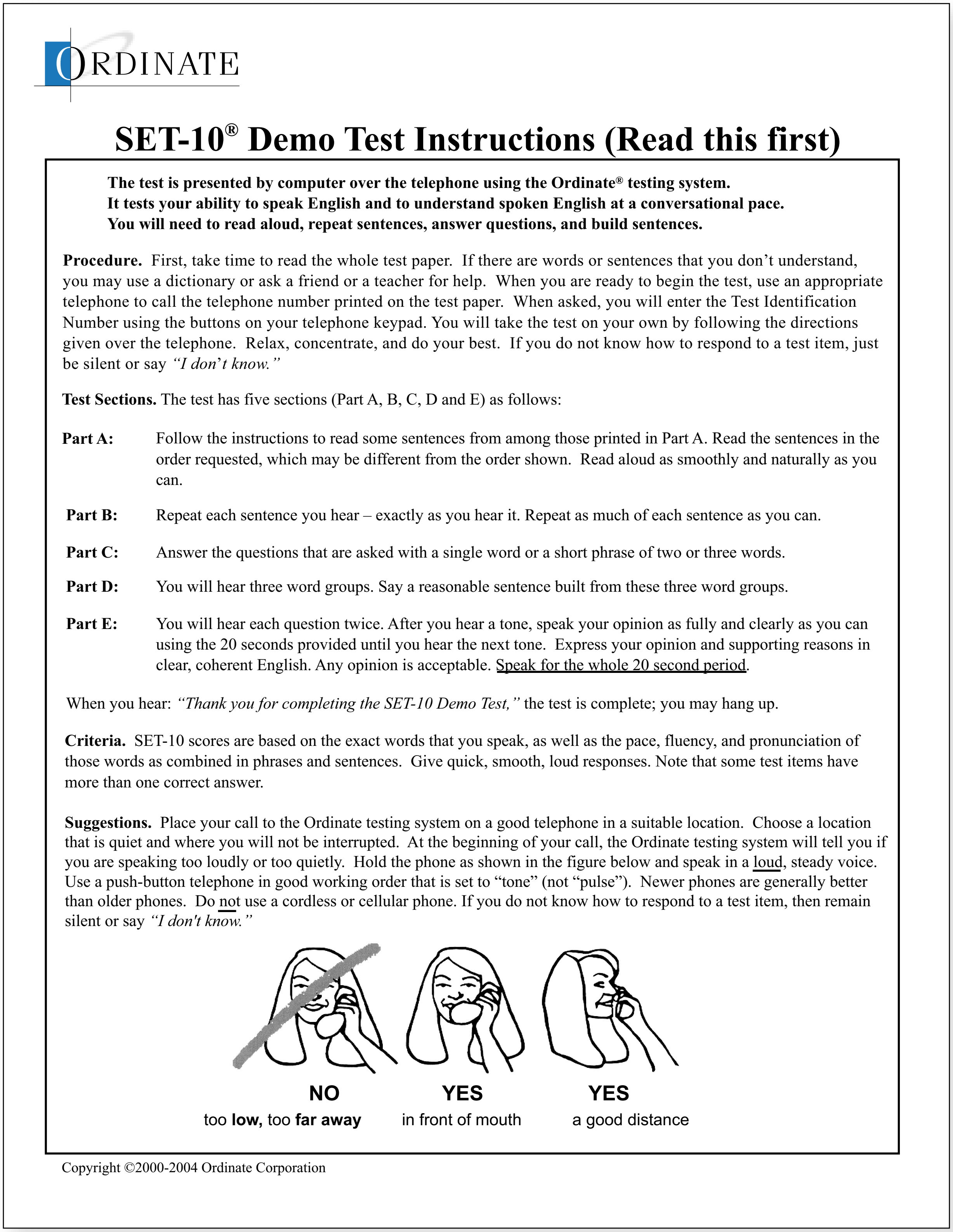
Schneider, W., & Shiffrin, R.M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84,* 1-66.

Van Turennout, M., Hagoort, P., & Brown, C. M. (1998). Brain Activity During Speaking: From Syntax to Phonology in 40 Milliseconds. *Science, 280,* 572-574.

Appendix A

Side 1 of the Test Paper: Instructions and general introduction to test procedures.

Note: These instructions are available in several different languages.



ORDINATE

SET-10® Demo Test Instructions (Read this first)

The test is presented by computer over the telephone using the Ordinate® testing system.
It tests your ability to speak English and to understand spoken English at a conversational pace.
You will need to read aloud, repeat sentences, answer questions, and build sentences.

Procedure. First, take time to read the whole test paper. If there are words or sentences that you don't understand, you may use a dictionary or ask a friend or a teacher for help. When you are ready to begin the test, use an appropriate telephone to call the telephone number printed on the test paper. When asked, you will enter the Test Identification Number using the buttons on your telephone keypad. You will take the test on your own by following the directions given over the telephone. Relax, concentrate, and do your best. If you do not know how to respond to a test item, just be silent or say *"I don't know."*

Test Sections. The test has five sections (Part A, B, C, D and E) as follows:

Part A: Follow the instructions to read some sentences from among those printed in Part A. Read the sentences in the order requested, which may be different from the order shown. Read aloud as smoothly and naturally as you can.

Part B: Repeat each sentence you hear – exactly as you hear it. Repeat as much of each sentence as you can.

Part C: Answer the questions that are asked with a single word or a short phrase of two or three words.

Part D: You will hear three word groups. Say a reasonable sentence built from these three word groups.

Part E: You will hear each question twice. After you hear a tone, speak your opinion as fully and clearly as you can using the 20 seconds provided until you hear the next tone. Express your opinion and supporting reasons in clear, coherent English. Any opinion is acceptable. Speak for the whole 20 second period.

When you hear: *"Thank you for completing the SET-10 Demo Test,"* the test is complete; you may hang up.

Criteria. SET-10 scores are based on the exact words that you speak, as well as the pace, fluency, and pronunciation of those words as combined in phrases and sentences. Give quick, smooth, loud responses. Note that some test items have more than one correct answer.

Suggestions. Place your call to the Ordinate testing system on a good telephone in a suitable location. Choose a location that is quiet and where you will not be interrupted. At the beginning of your call, the Ordinate testing system will tell you if you are speaking too loudly or too quietly. Hold the phone as shown in the figure below and speak in a loud, steady voice. Use a push-button telephone in good working order that is set to "tone" (not "pulse"). Newer phones are generally better than older phones. Do not use a cordless or cellular phone. If you do not know how to respond to a test item, then remain silent or say *"I don't know."*

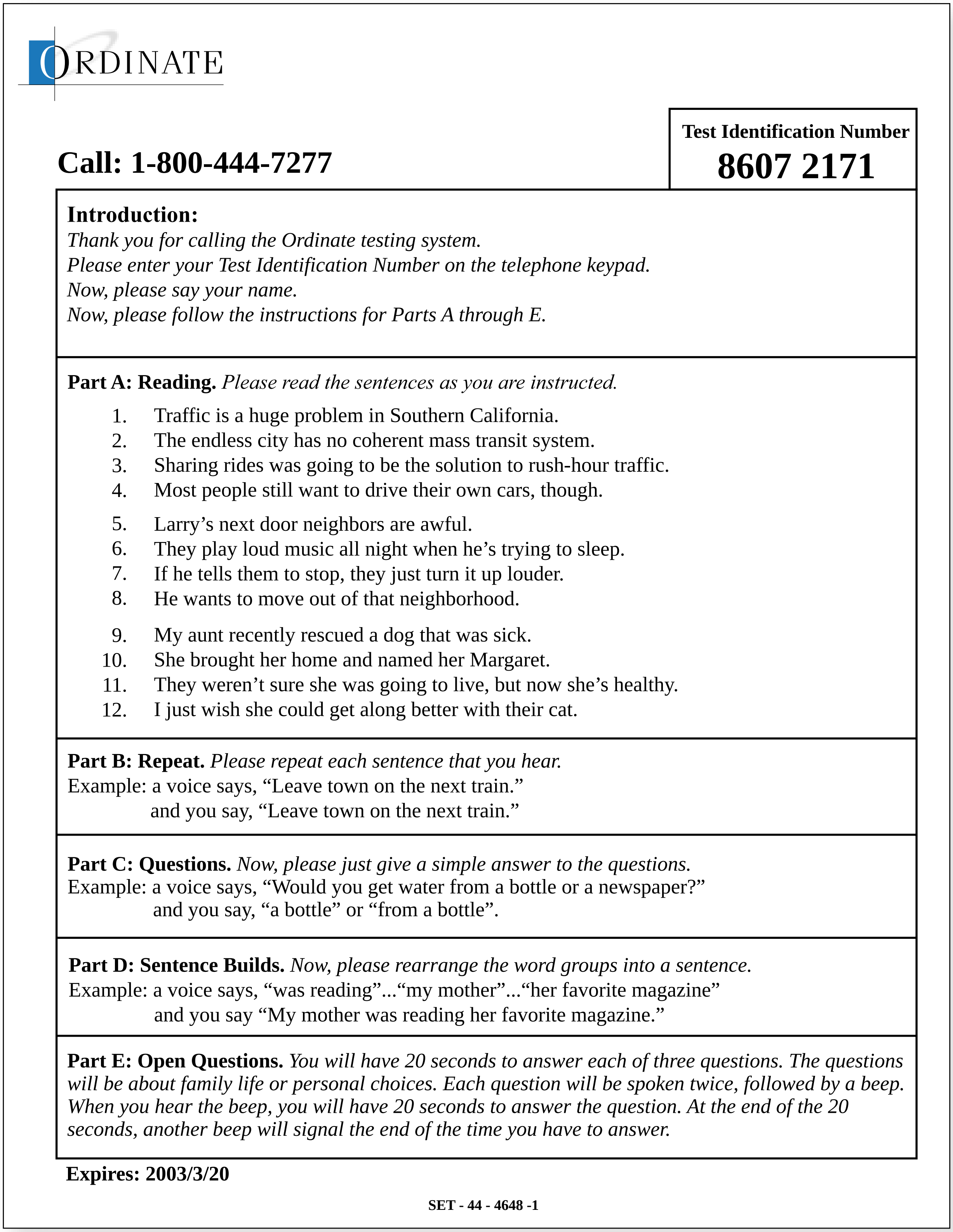
NO
too **low,** too **far away**

YES
in front of mouth

YES
a good distance

Appendix B

Individualized test form (unique for each test taker) showing Test Identification Number, Part A: sentences to read, and examples for all sections.



ORDINATE

Test Identification Number
8607 2171

Call: 1-800-444-7277

Introduction:
Thank you for calling the Ordinate testing system.
Please enter your Test Identification Number on the telephone keypad.
Now, please say your name.
Now, please follow the instructions for Parts A through E.

Part A: Reading. *Please read the sentences as you are instructed.*

1. Traffic is a huge problem in Southern California.
2. The endless city has no coherent mass transit system.
3. Sharing rides was going to be the solution to rush-hour traffic.
4. Most people still want to drive their own cars, though.
5. Larry’s next door neighbors are awful.
6. They play loud music all night when he’s trying to sleep.
7. If he tells them to stop, they just turn it up louder.
8. He wants to move out of that neighborhood.
9. My aunt recently rescued a dog that was sick.
10. She brought her home and named her Margaret.
11. They weren’t sure she was going to live, but now she’s healthy.
12. I just wish she could get along better with their cat.

Part B: Repeat. *Please repeat each sentence that you hear.*
Example: a voice says, “Leave town on the next train.”
and you say, “Leave town on the next train.”

Part C: Questions. *Now, please just give a simple answer to the questions.*
Example: a voice says, “Would you get water from a bottle or a newspaper?”
and you say, “a bottle” or “from a bottle”.

Part D: Sentence Builds. *Now, please rearrange the word groups into a sentence.*
Example: a voice says, “was reading”...“my mother”...“her favorite magazine”
and you say “My mother was reading her favorite magazine.”

Part E: Open Questions. *You will have 20 seconds to answer each of three questions. The questions will be about family life or personal choices. Each question will be spoken twice, followed by a beep. When you hear the beep, you will have 20 seconds to answer the question. At the end of the 20 seconds, another beep will signal the end of the time you have to answer.*

Expires: 2003/3/20

SET - 44 - 4648 -1

ORDINATE

1040 Noel Drive
Menlo Park, CA 94025 USA
phone +1.650.327.4449
fax +1.650.328.8866
www.ordinate.com