# Automatically Scored Spoken Language Tests for Air Traffic Controllers and Pilots

*Jennifer Balogh, Jared Bernstein, Masa Suzuki, Matthew Lennig*
Ordinate Corporation, Menlo Park, California, USA

## Abstract

Ordinate Corporation has been in the business of developing automatically scored spoken language tests since 1996. Ordinate currently offers automatically scored tests for English, Spanish and Dutch. These tests are being used by many governments, corporations and educational institutions around the world.

Ordinate's automatically scored tests have significant administrative benefits over human scored tests since they can be administered from any standard phone at any time and results are available within a few minutes of the test administration. In addition, Ordinate's tests go through a rigorous validation process to ensure they are as accurate and reliable as the gold standard human rated interview type test.

In December 2005, Ordinate entered into a CRDA with the FAA Academy to develop a new test, the *Versant Aviation English Test,* designed specifically for the aviation industry. The tasks and items types of the *Versant Aviation English Test* reflect the specific requirements of the International Civil Aviation Organization (ICAO) aviation English standard.
In this paper, we will describe the test, discuss how it is criterion-referenced to the ICAO standard, and review how its reliability and validity will be established.

## Introduction

The International Civil Aviation Organization (ICAO) set a mandate that air traffic controllers and flight crews need to attain a minimum English proficiency level for safety reasons by March 2008. The minimum English proficiency level is defined as Level 4 on the ICAO proficiency scale. In this time frame, air traffic controllers and pilots are required to demonstrate the ability to speak and understand aviation radiotelephony communications in English as well as common English related to the aviation domain (ICAO, 2004).

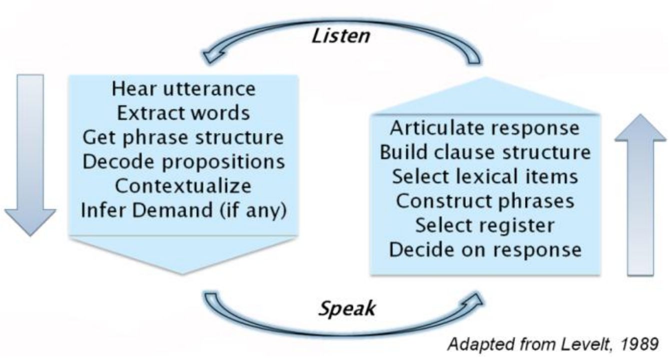
Over the past decade, advances in speech recognition technology have enabled the development of an automatically administered and scored spoken language tests in English (Bernstein, De Jong, Pisoni, Townshend, 2000). During the test, human-recorded prompts are played over a land-line telephone, and test-takers' responses are automatically scored using speech recognition and other computerized speech processing technologies. Because the test is automated, large numbers of tests can be administered and scored rapidly, maintaining consistently high accuracy. These English tests, as well as tests of spoken Spanish and Dutch are developed, administered and scored by Ordinate Corporation, a wholly-owned subsidiary of Harcourt Assessment, Inc.

In order to respond to the need in the aviation industry for a reliable, valid, and efficient test that can meet ICAO English language proficiency requirements, Ordinate is currently working with the FAA and professors at Oklahoma State University to develop an automated spoken English test for Aviation called the *Versant Aviation English Test* that can assess a test-taker's English language proficiency based on the ICAO English language proficiency scale.

In what follows, we describe the construct of the test, the test administration and structure, the computer architecture, and the planned validation process. Results from Ordinate's existing English test, *Versant for English,* will be provided to show how tests built on Ordinate's test system typically perform.

## Test Construct

The *Versant Aviation English Test* is intended to measure the *facility in spoken aviation English and common English in the aviation domain*. This is the ability to understand spoken English both within the aviation radiotelephony phraseology and on topics related to aviation (such as movement, position, time, duration, weather, animals, etc.), and to be able to respond appropriately in intelligible English at a native-like conversational pace. This definition relates to what occurs during the course of a spoken conversation as schematized in Figure 1, adapted from Levelt (1989).



Adapted from Levelt, 1989

Figure 1. Conversational processing components in listening and speaking.

Spoken language facility is essential to successful oral communication – if language users cannot track what is being said, extract meaning as speech continues, and then formulate and produce a relevant and intelligible response in real time, they will not be able to interact effectively in operational communication. In the administration of the *Versant Aviation English Test,* the Ordinate testing system presents a series of discrete prompts to the test-taker over the telephone at a native conversational pace. These integrated "listen-then-speak" items require real-time receptive and productive processing of spoken language forms. Because the *Versant Aviation English Test* requires real-time language processing, it measures the degree of automaticity in basic encoding and decoding of oral language. Automaticity is 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). Automaticity is required in order for the speaker/listener to pay full attention to actions and to what needs to be said rather than thinking through how the spoken message should be structured. Automaticity is critical for non-native English speakers since aviation professionals need to be able to handle complicated and unexpected turns of events without having to consciously attend to the process of understanding and producing the language.

## Arguments Against Automation

Before tests are administered, a unique test paper is created for each test-taker. The test paper presents the telephone number to call, a unique Test Identification Number (TIN), written instructions for each of the tasks along with an example, and printed test items for specific tasks. For test administration, the test-taker calls a phone number printed on the test paper and enters the TIN using the telephone keypad. The test is then administered automatically over the phone. The *Versant Aviation English Test* will take about 30 minutes to complete.

The Test Identification Number ensures that each test-taker is presented with a different set of test items. Test items are drawn from large item pools in a stratified random fashion, such that within each subpart of the test, the easier items are presented earlier, while more difficult items are presented later.

A score report is generated a few minutes after test completion and will be posted on Ordinate's website. The score report will report the six subskills that ICAO specifies: Pronunciation, Structure, Vocabulary, Comprehension, Fluency, and Interactions. Score users can view and download scores from Ordinate's password-protected secure website. The website will allow score users to listen to select spoken responses from test-takers.

## Test Structure

The specific tasks in the test were designed to provide information for one or more of the measurement subcomponents: Pronunciation, Structure, Vocabulary, Fluency, Comprehension and Interactions.

Responses longer than a single word or short phrase can be used to estimate Pronunciation. For this reason, all sentence-level responses from the test will be used for Pronunciation scoring. Ordinate's system extracts information about the stress and segmental forms of the words in the responses and the pronunciation of the segments in the words within their lexical and phrasal context. A Reading task begins the test to collect an additional sample of the test-taker's speech. In the Reading task, test-takers read printed, numbered sentences, one at a time, in the order requested. Reading items are grouped into sets of four sequentially coherent sentences. One group is in common, everyday English and the other reflects aviation phraseology. The sentences are presented on a printed test paper that the test-taker receives before taking the test.

For the Structure subcomponent, the goal of the test is to measure the ability to parse syntactic structure, manipulate linguistic units, and produce complex sentences. In a task section designed to assess these skills called Repeat, the test-taker listens to a sentence and then tries to repeat it verbatim. An example sentence in English is "My flight leaves on Saturday." 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 less familiar with the spoken language. Highly proficient speakers can generally repeat sentences that contain many more than seven syllables because these speakers are very familiar with the language's phrase structures and common syntactic forms. If a person habitually processes five-word phrases as a unit (e.g. "the very large descending aircraft"), 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. The Repeats will include both basic and complex grammatical structures and sentence patterns.

The other task section that extracts information about Structure is called Readbacks. In this task, test-takers are presented with a radiotelephony message and are asked to give an appropriate readback to confirm their understanding of the message. The test-taker is expected to produce a readback using

ICAO phraseology. Readbacks measure the Structure subcomponent in predictable, work-related language.

For Vocabulary, test-takers are asked to listen to an orally presented story or incident that deals with an aviation-specific topic and then describe what happened in their own words. Test-takers must identify words in phonological and syntactic context, extract propositions, and then paraphrase what was said. Scoring of the Story Retellings will focus on the test-taker's vocabulary range as opposed to grammatical structure, which is addressed in other task items.

For Fluency scores, constrained responses longer than a single word or short phrase are used to assess Fluency. Although the same responses used to estimate Pronunciation ability are used here, the scoring is independent. For Fluency, features such as rate of speaking and the position and length of pauses are analyzed.

For the Comprehension subcomponent, the test needs to assess the test-taker's ability to understand common and work-related words and concepts in sentence context. To do this, the test presents the test-taker with a series of questions that can be answered with a word or short phrase. An example of a Short-Answer Question is, "Would you get water from a bottle or a newspaper?" The question requires the test-taker to identify the lexical items of the question in phonological and syntactic context, infer the demand proposition, and then say an appropriate response. Lexical items are based on the ICAO list of priority lexical domains including topics such as animals, numbers, movement, time, transportation, and weather. Since items are recorded in different native and non-native voices, the test-taker must be able to comprehend a range of speech varieties.

Finally, two different aspects of speech will contribute to the Interactions subscore: the content of what is said and the test-taker's response time. In a task called Corrections and Confirmations, the test-taker will hear a radiotelephony message, either from the air traffic controller's perspective or the pilot's perspective. As additional support, the text will be printed on the test paper. Then, the test-taker will hear a readback, which might contain the correct information, wrong information, or a request for more information. The test-taker is expected to respond to the message appropriately using ICAO phraseology. For example, if the response contains wrong information, the test-taker is expected to provide correct information. Some items will reflect routine communications/situations; others will cover less routine communications/situations, and a small proportion will explore unexpected communications/situations. The rubrics in the ICAO manual also include response time as an important aspect of the Interactions scale. Ordinate's testing system incorporates a measurement of the initial latency of the test-taker's response which will contribute to the Interactions subscore.

Table 1 lists the eight tasks as they appear in the *Versant Aviation English Test.*

Table 1. Tasks in the *Versant Aviation English Test*

| Tasks |
|---|
| Part A. Reading – Aviation |
| Part B. Reading – Common English |
| Part C. Repeat |
| Part D. Short Questions |
| Part E. Readbacks |
| Part F. Corrections and Confirmations |
| Part G. Story Retellings |

The tasks in the *Versant Aviation English Test* provide direct measures of the test-taker's listening and speaking ability. The ICAO manual states that tests that only assess phraseologies or that test only plain English are not appropriate. The *Versant Aviation English Test* addresses this requirement by including both aviation specific-phraseology tasks and common English tasks. The tests are administered on Ordinate's computerized testing system described below.

## Ordinate's Testing System

The system architecture that handles test administration consists of a telephone network, telephony equipment that connects and disconnects the calls, and servers that present audio prompts and capture the test-taker's responses, as shown in Figure 2.

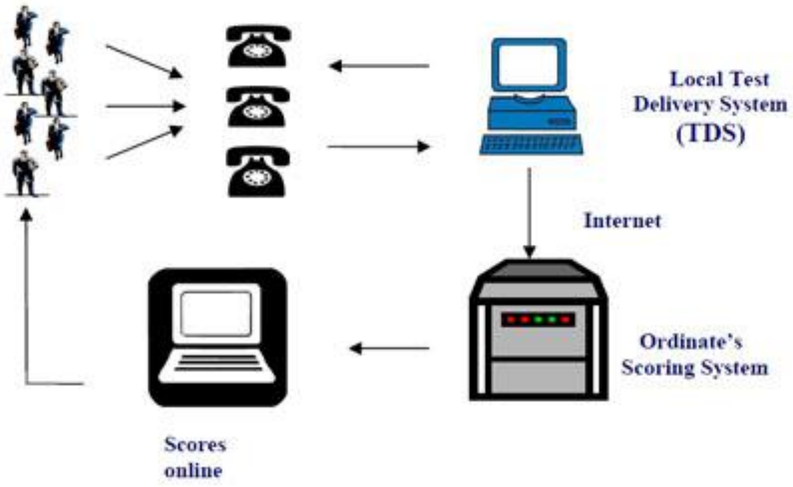


Figure 2. Information flow in Ordinate's test operation.

When a test-taker calls into the system, the call is connected at a Test Delivery Server (TDS). Countries like Japan, Korea, China, and the Netherlands have a local TDS, so that test-takers in different countries can call a local toll-free number. The TDS assembles the appropriate test items, as identified by the unique TIN on the test paper to the TDS. The TDS then delivers the items over the telephone. The local TDS communicates with Ordinate's scoring system via a Virtual Private Network (VPN). Ordinate's scoring system has a database that stores all the tests' human-recorded audio files for the items as well as all the instruction prompts. The database also contains information about each item, such as its level of difficulty, and information about the test design, including the number of items presented in each task section.

The system listens to the test-taker's response and then automatically proceeds to the next item. Responses travel from the phone to the local TDS to the scoring system. As the system receives responses, it starts extracting information from the utterance for scoring. The system uses automatic speech recognition, speech to text alignment, and non-linear models to perform automatic scoring. The spoken response files sent to the Ordinate testing system are recognized and scored by the automated speech recognition system and other automated computerized scoring systems.

Ordinate uses an HMM-based (Hidden Markov Model) speech recognition system. The acoustic models of the speech recognition system, or representations of the sounds or *phones* produced when speaking English, are trained on thousands of non-native English speakers. Non-native speakers are used so that the system can accurately extract words from heavily accented speech.

This speech recognizer uses a dictionary, which lists the common pronunciations for each word that the system should recognize. A language model, or a representation of the sequence of words the speaker is likely to say, is also developed. For example, if the respondent is asked to repeat the sentence, "My flight leaves on Saturday," then it is very likely that the test-taker will say the words "My flight leaves on Saturday." The high probability associated with this string of words is encoded in the language model for this sentence. The language models not only contain the most likely strings of words that test-takers are expected to say, but also the types of mistakes and disfluencies that non-native English speaking air traffic controllers and pilots are most likely to make.

Using the acoustic models, dictionary, and language models, the speech recognition system employs statistical methods to identify the string of words that best matches the respondent's speech. The hypothesis of what the respondent said is then compared to the words in the item. Models based on Item Response Theory (IRT) use the 'correctness' of the content of individual responses in addition to the item's difficulty level to produce estimates of the test-taker's Vocabulary, Comprehension, Structure, and Interactions abilities.

Other information is also extracted from the respondent's utterance such as speaking time, rate of speech, and mean pause duration. These and other paralinguistic parameters are then input into non-linear models that are optimized to predict how human listeners would judge the responses with regard to pronunciation and fluency.

Scoring algorithms then compute final scores that are posted to a secure web site a few minutes after the test is completed. Of interest to the test score user is the test's validity as discussed in the next section.

## Validation Process

The general approach to validation highlights three metrics as evidence of the test's quality: high reliability, effective separation between appropriate samples of test-takers, and a strong correlation with human scores. Successfully achieving these metrics relies not only on the integrity of the test, but also on a rigorous methodology.

Data from both native and non-native test-takers will be used for the validation process. Usually, data collected during the development process is set aside for this purpose. Native speakers will be professionals in the field who live in a variety of regions and countries and represent a range of age groups. Non-native speakers will be air traffic controllers and pilots with a broad range of English proficiency levels and native languages.

### *Reliability*

From the validation data collection, split-half reliability will be computed for all subscores and for the Overall score of the test. These reliabilities will use the Spearman-Brown Prophecy Formula to correct for split-half underestimation. For the English production-level test built on top of the same testing framework, Overall score reliability is .97. For Ordinate's Spoken Spanish Test, the reliability is 0.96.

### *Native and Non-native Group Performance*

An indicator of construct validity will be the separation of scores of native and non-native aviation professionals. The assumption is that native English speakers as a group possess a high degree of facility in spoken English. Therefore, if natives who are professionals in the aviation industry obtain high scores while non-native speakers are distributed over a wide range of scores, then the expected distinction between the groups lends support of the test's validity.  Although a significant separation between the groups is expected, the pattern of results will most likely be different from Ordinate's general *Versant for English* test.  For *Versant for English,* learners of English as a second or foreign language are distributed over a wide range of scores, many of them at the very low end of the scale.  However, fewer non-native aviation professionals are expected to score low on the *Versant Aviation English Test* given the ICAO English language standards.  Even so, the test will have high discriminatory power among English language learners in the aviation industry.

### *Correlations between Test Scores and Human Ratings*

The third validity metric is the correlation between the test scores and human scores.  Responses will be transcribed by professional human transcribers and rated independently by human experts for pronunciation and fluency.  In this way, a sample of calls will be scored without automatic speech processing technologies.  These human-generated scores will be compared with the machine-generated scores to determine the accuracy of the automatic scoring.

Similar analyses were performed for the general *Versant for English.* Table 2 below presents correlations between the human-generated and machine-generated scores.

Table 2. Correlations of human-generated and machine-generated scores.

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

The correlations in Table 2 suggest that the machine-generated scores for *Versant for English* systematically correspond with human-generated scores.  Similar results are expected for the *Versant Aviation English Test* given the similarity of the test architecture and computerized scoring system.

## Conclusion

The *Versant Aviation English Test* will be an automatically administered and scored test of spoken English for the aviation industry. The benefit of this type of test, compared to human-conducted and scored interviews, is that it can be administered to large numbers of test-takers, scored rapidly, yet maintain high reliability and consistent operational validity. The test architecture is designed such that different tasks provide information both about the content of what is said and the manner in which the responses are spoken.  The computerized testing system provides a general means of automatically administering and scoring tests, with language models trained specifically from aviation professionals who are non-native speakers of English. Finally, we expect that the validation process will show that the test is both reliable and valid.  Data from Ordinate's *Versant for English* test show that tests built on top of Ordinate's testing framework are reliable, can distinguish high proficiency from low proficiency groups of test-takers, and correlate highly with human scores.  Similar results are expected for the *Versant Aviation English Test.*

## 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*. University of Abertay Dundee, Scotland, 57-61.

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

International Civil Aviation Organization (2004). *Manual on the implementation of ICAO language proficiency requirements, First Edition.*

Jescheniak, J.D., Hahne, A., & Schriefers, H.J. (2003). Interformation 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.