

Automatically Scored Spoken Language Tests

Mr. Larry Rothenberg, Dr. Jennifer Balogh, and Dr. Jared Bernstein
Harcourt Assessment
Menlo Park, CA, USA

The *Versant™ Aviation English Test* (VAET), which uses Ordinate technology, is a speaking and listening test that is designed to evaluate spoken English in the aviation domain. As an automatically administered and automatically scored test, the VAET will establish a benchmark for operational accuracy and consistency. In this paper, we describe the test and discuss how it meets the specific requirements of the International Civil Aviation Organization (ICAO) standard as described in the *Manual on Implementation of the ICAO Language Proficiency Requirements, First Edition, Document 9835* (referred to below as the ICAO Manual).

Test purpose The *Versant Aviation English Test* (VAET) measures facility in spoken aviation English and common English in the aviation domain; that is, the ability to understand spoken English with regard to aviation radiotelephony phraseology and on topics related to aviation (such as movement, position, time, weather), and the ability to respond appropriately in intelligible English at a fully-functional pace.

The VAET is based on a psycholinguistic view of language competence, which emphasizes the development of language skills and the real-time processes that underlie the performance of these skills. Eysenck & Keane (1995) review studies that show that complex cognitive skills that are used in language become automatic in skilled performance, and therefore do not absorb any of the attentional capacity of the speaker-hearers. A corollary of a psycholinguistic view relevant to testing is that if a speaker must devote a large amount of attention to how to say something, i.e., by having to consciously select words and syntactic structure, then the speaker will not be able to pay as much attention to the content of what is said; therefore the measurement of fluent, automatic control of core language will predict the complexity of content that can be produced or understood during the rapid pace of spoken dialogue. 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 in effective radiotelephony communication, especially when attention is focused on aviation-related tasks at hand. The VAET does not test *knowledge about* the language, but rather tests the test taker's *facility with* the language. In this way, the VAET directly evaluates speaking and listening through an audio exchange, as deemed appropriate by the ICAO Manual.

Test Development Team and Development Process

The ICAO Manual offers specific recommendations regarding the training and background of the professionals involved in developing a language test for aviation professionals. The VAET test development team consisted of numerous language testing experts and relevant stakeholders in aviation language testing. Test Developers comprised a university professor with a Ph.D. in Linguistics, specializing in applied linguistics and language testing; a founder of a test development company with a Ph.D. in Psycholinguistics and over 10 years of experience developing and conducting research on second/foreign language tests; a manager of a test development group with a Ph.D. in Psychology and three years of experience developing and conducting research on second/foreign language tests; and a professional test developer with an M.A. in TESOL and over four years of experience developing and conducting research on second/foreign language testing. Another university professor of English, specializing in oral proficiency testing and aviation English also consulted on the project. Many aviation professionals were involved to ensure accuracy of content and to provide expert ratings used to develop automatic scoring models. Aviation professionals included a program manager for aviation training and an ICAO committee member with 25 years of experience in resource development and language training;

a consultant for the ICAO English language competency project with over 30 years of experience as an air traffic control officer in the UK and Australia; a professional pilot with over 10 years of experience as a pilot in the UK; an international air traffic training program manager with over 10 years of experience in international air traffic training; and a program officer for a government English language center and over 25 years of experience in curriculum development for aviation English language training. The recorded voices used in the test include two retired air traffic controllers, two active air traffic controllers, one retired commercial pilot, and two active commercial pilots. In addition, over 1200 professional air traffic controllers and pilots from almost 90 different countries representing approximately 50 languages participated in the VAET's trial and validation research studies.

The development of the VAET assessment followed a rigorous process that entailed clearly defining the test construct, creating a detailed test specification, writing an item specification to be used by item writers, developing the items, coordinating review of the items by trained professionals, revising the items, creating a data collection system, collecting feedback on the test administration from stakeholders and revising the test instructions, trialing the items with both native and non-native English speaking aviation professionals from a variety of countries, analyzing the data, making a final selection of items, enabling automatic scoring of the items, and validating the test. The process aligns with the test development process recommended by the ICAO Manual (6.5.4).

Test Design

The VAET testing procedure consists of eight tasks as described in Table 1.

Test Section	Description	Type of English	Measured Skill Areas
A: Reading Aviation Phraseology	Test taker is asked to read sentences aloud, one at a time.	ICAO Phraseology	Pronunciation Fluency
B: Reading Common English	Test taker is asked to read sentences aloud, one at a time.	Common English	Pronunciation Fluency
C: Repeat	Test taker listens to a sentence in common English and tries to repeat it verbatim.	Common English	Structure Pronunciation Fluency
D: Short Answer Questions	Test taker answers questions with single words or short phrases.	Common English	Comprehension Vocabulary
E: Readbacks	Test taker listens to a radiotelephony message and says a readback.	ICAO Phraseology	Structure Pronunciation Fluency
F: Confirmation and Correction	Test taker hears a short dialogue between a pilot and air traffic controller and responds appropriately.	ICAO Phraseology, Common English	Interactions
G: Story Retellings	Test taker hears about an incident or scenario and tries to retell it in his or her own words.	Common English	Vocabulary
H: Open Questions	Test taker hears a question and explains his or her answer.	Common English	For validation

Table 1. Tasks in the VAET.

In the VAET, both ICAO phraseology and common English appear throughout the test. For example, the Repeats section presents common English sentences with aviation content, while the Readbacks section uses ICAO phraseology. In the Corrections and Confirmation section, ICAO phraseology is used except

in those cases in which the test taker is presented with a non-routine situation and must use common English to clarify, correct, or confirm information. The test follows the recommended model in the ICAO Manual, in which a test is “comprised of a mix of both aviation-specific content alongside less aviation-specific content” (Section 6.8.7). Even with common English items, the content of the VAET was designed to address work-related topics in the aviation domain. For the common English tasks such as Repeats and Short Answer Questions, the ICAO list of priority lexical domains guided the choice of lexical items and topics. In addition, transcripts of conversations from actual pilots and air traffic controllers were used as reference material for creating items in the Readback, Confirmation and Correction, and Story Retellings sections of the test.

Each specific task in the test is designed to provide information about the test taker’s ability on one or more of the measurement subcomponents: Pronunciation, Structure, Vocabulary, Fluency, Comprehension and Interactions:

Pronunciation

Responses longer than a single word or short phrase are used to estimate Pronunciation. 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. These measures are scaled according to native and non-native distributions and then are rescaled such that they optimally predict human judgments of pronunciation.

Structure

For the Structure sub-skill, the goal is to measure the ability to control basic and complex grammatical structures. 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. As the sentences increase in length and complexity, the task becomes increasingly difficult for less proficient speakers who are less familiar with the language’s phrase structure and common syntactic forms. Generally, repetition of material is constrained by the size of the linguistic unit (e.g. “the very large descending aircraft”) that a person can process in an automatic or nearly automatic fashion. The Repeats 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.

Vocabulary

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. Some topics are designed to be common, concrete, and familiar while others are designed to be more abstract and unfamiliar. Test takers must identify words in continuous phonological and syntactic context, extract propositions, and then paraphrase what was said. Scoring of the Story Retellings focuses on the test taker’s vocabulary range and accuracy.

Fluency

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 for Fluency, the scoring is independent. For Fluency, features such as rate of speaking and the position and length of pauses are analyzed. The measures are scaled according to native and non-native distributions and then are rescaled such that they optimally predict human judgments of fluency based on the ICAO Rating Scale.

Comprehension

For the Comprehension subcomponent, the test assesses the test taker's ability to understand common, work-related words and concepts in sentence context. To do this, the test taker is presented with a series of questions that can be answered with a word or short phrase. An example of a Short-Answer Question is, "<http://renoir.ordinate.com/OMI/viewCue?cue=31114>Is land that's almost entirely surrounded by water a peninsula or a pond?" 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 speak 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. Because items are recorded in different native and non-native voices, the test taker must be able to comprehend a range of speech varieties.

Interactions

In a task called Corrections and Confirmations, the test taker hears a radiotelephony message, either from an air traffic controller's perspective or a pilot's perspective. Then, the test taker hears a response, 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 when possible. If the response contains wrong information, the test taker is expected to provide correct information. Some items reflect routine communications/situations; others cover less routine communications/situations; and a small proportion explore unexpected communications/situations.

The ICAO rubrics highlight two main dimensions for assessing *Interactions*: the speed of the response and how appropriate the response is. For response latency, different levels provide different descriptions of response speed. At the operational level, for example, responses are "immediate" whereas at the elementary level, "response time is slow." In interviews, these judgments would be made from subjective impressions. In contrast, the Ordinate testing system records precisely how many milliseconds it takes the test taker to respond to an item. Whereas human-based interviews may involve biased perceptions, the machine records an objective measure of response time. Latency measurements are compared to the same measurements from proficient speakers and are then rescaled. The calibrations are done the same way for each test taker, so there is no inconsistency or bias.

To identify what the test taker has said and judge the appropriateness of the content of the response, speech processing technologies recognize what the test taker has said, and then the Ordinate system uses Item Response Theory (IRT) to produce an estimate of the test taker's ability according to the item's difficulty. The analysis is not based on just one rater's impression of what is appropriate, but rather comparisons of the test taker's response to hundreds of responses from native and non-native pilots and air traffic controllers of all proficiency levels. These comparisons help to precisely pin-point the test taker's level of ability through robust statistical analyses.

It may not be obvious that a computer can evaluate the Interactions subskill. It may seem that a multi-step dialogue is essential for its assessment. The ICAO Manual, however, outlines two different testing procedures: direct and semi-direct testing. In direct tests, the test taker is presented with elicitation prompts from a tester while being observed by one or more raters. In the semi-direct testing procedure, pre-recorded prompts are presented to the test taker and responses are usually recorded and analyzed later. The semi-direct approach provides information equivalent to a multi-step dialogue. The VAET uses a semi-direct administration procedure. Recorded prompts from a variety of voices are presented to the test taker over the telephone. Each prompt elicits a response from the test taker that is recorded digitally and then analyzed. As described in the ICAO Manual, both procedures directly assess the test taker's language speaking or listening ability and are appropriate for assessing language proficiency (6.7.10). The fact that a computer can measure latency of speech and can evaluate whether or not a response is appropriate (in both predictable and unexpected situations) indicates that a computer can, in fact, measure ability on the *Interactions* scale.

Together, the tasks in VAET provide direct measures of the test taker's listening and speaking ability, according to the six subskills in the ICAO Rating Scale.

Test Security

For test security, the VAET is not made up of static test forms, but of dynamically generated forms that are unique from one test administration to the next. Each test is created from a structured random selection of items from a large item pool, limiting item exposure. The tests are equivalent in difficulty because the system takes into consideration item difficulty as it dynamically assembles test forms. Harcourt endorses the use of VAET scores for making valid decisions about spoken English skills of aviation professionals, 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 VAET themselves, including photo identification checks and testing supervision, or by having tests administered by trusted third parties who follow the same procedures. As is recommended in section 6.7.9 of the ICAO Manual, each test performance is audio-recorded and can be referenced for verification and record-keeping purposes.

Validation

The general approach to validation of the VAET highlights two important metrics of the test's quality: high reliability in operation and evidence of validity.

Reliability

The reliability of a test refers to the precision, consistency and stability of test scores. One estimate of score consistency is split-half reliability. Split-half reliability is calculated from data on a single administration of the test. Scores from equivalent halves of each subtest representing approximately parallel forms are correlated. The correlation is corrected for split-half underestimation by using the Spearman-Brown Prophecy Formula. Split-half reliability will be computed for all subscores. The higher the reliability coefficient, the greater confidence one can place on the consistency and precision of the scores. For the English and Spanish production-level tests built within the same framework, Overall score reliabilities are .97 and .96 respectively. When VAET is launched, the split-half reliability as observed in operation will be reported. Results are expected to be similar to the current production tests. Because all the VAET tests are scored by the same machine using the same algorithms, scores are not affected by differences in rater severity or bias. Consistency and objectivity in machine scoring eliminate this source of measurement error. Because of this, the VAET will establish a benchmark of unbiased, consistent performance rating across countries.

Content Validity

Content validity is the degree to which the test items represent the content that the test is designed to measure. Content-related evidence of validity was provided by both expert judgment and empirical item analysis. Each item was reviewed by subject matter experts to ensure content relevancy, conformity to common usage, and compliance with ICAO standards. In addition, each item was analyzed statistically using data from both native and non-native speakers of English. If an item was not answered correctly by a certain percentage of a sample of natives, then it was unclear if the item was measuring spoken English or another underlying ability, and the item was not included in the test. Other statistical analyses were performed to ensure that each item was effective at discriminating language ability.

Construct Validity

Construct validity is the extent to which test scores can be interpreted as a measure of the intended construct. An indicator of construct validity is the separation of native and non-native aviation professionals with regard to Overall scores. The assumption is that native English speakers as a group possess a high degree of facility in spoken English. In contrast, non-native speakers have different levels of English language proficiency, ranging from very little facility to native-like facility. Therefore, if native English-speaking pilots and air traffic controllers obtain high scores while non-native pilots and air traffic controllers are distributed over a wide range of scores (including the critical Operational Level 4), then the expected distinction between the groups lends support to the test's validity. This analysis will be reported when the VAET is launched.

Concurrent Validity

Another validity metric is the correlation between the test scores and other well-established measures of oral language proficiency. Each test taker's proficiency level was estimated by having human experts rate responses to Open Questions and Story Retellings using the ICAO rubrics. These ratings will be correlated to test scores and reported when the VAET is launched.

Taken together, the reliability analyses and the three lines of validity evidence will indicate the test's level of consistency and the fact that test scores can be interpreted as a measure of spoken English language ability in the aviation context.

Conclusion

The Manual on the Implementation of ICAO Language Proficiency Requirements (2004) offers standards for testing language proficiency of air traffic control personnel and flight crews involved in flight operations. The *Versant Aviation English Test (VAET)* is a speaking and listening test designed to meet the ICAO standards. The test is taken over the telephone and is administered and scored automatically. All test taker responses are recorded and stored in Ordinate's secure testing system. VAET is intended to measure facility in spoken aviation English and common English in the aviation domain. As such, it is designed to test facility with spoken English as opposed to knowledge about English. The test development team consisted of experienced aviation professionals as well as language testing experts with appropriate degrees and many years of experience in language testing. The test tasks present a mixture of items with ICAO phraseology and common English addressing aviation work-related topics and are designed to measure all six language skills (Pronunciation, Structure, Vocabulary, Fluency, Comprehension, and Interaction) independently. At the launch, validation data will be presented including the reliability and evidence of content, construct, and concurrent validity.

In addition to meeting the standards outlined in the ICAO Manual, the automated nature of VAET provides many benefits including consistent administration, the ability to handle large volumes of test takers, unbiased and consistent scoring, and immediate results. Given that English language proficiency needs to be assessed across many different countries, having an unbiased, consistent standard of measurement will most likely prove to be an important factor in English language assessment for aviation.

References

Eysenck, M. & Keane, M. (1995). *Cognitive Psychology*. Hove, UK: Psychology Press.
International Civil Aviation Organization, (2004). Manual on the implementation of the ICAO language proficiency requirements, First Edition, Document 9835.

— — — — —