

A Case for Automation in Aviation English Language Assessment

Jennifer Balogh, Ph.D., Pearson Education, Inc.

Introduction

As the International Civil Aviation Organization (ICAO) adopts English language proficiency requirements for radiotelephony communications, international air carriers and government authorities are preparing their pilots and air traffic controllers to meet these ICAO standards. As a part of this effort, companies, organizations, and governments are in the process of making careful decisions about how to evaluate English language proficiency in the aviation context.

Traditionally, spoken language ability has been assessed with Oral Proficiency Interviews (OPIs). In an OPI, one or more trained raters ask the examinee questions and elicit a spoken performance. The rater(s) then analyze these responses during the course of the interview, often adjusting the complexity of the questions to meet the perceived proficiency level of the examinee. The raters assign the examinee a performance level according to how closely the examinee's performance matches predefined descriptors for that level.

Although OPIs are a common test of proficiency, they are not the only approach to the evaluation of spoken language. The past decade has introduced significant changes to the field of spoken language assessment due to advances in automatic speech recognition and other speech processing technologies. These new technologies have enabled the development of automatically administered and scored spoken language tests (Bernstein, De Jong, Pisoni, Townshend, 2000). In response to the ICAO English language standards, a new test of aviation English is currently

being developed that uses the same technologies. The excitement surrounding the introduction of such a spoken language test and what it could mean for large-scale assessments does not come without some hesitancy. There are three main areas of concern specific to the aviation sector and the implementation of the new ICAO regulations: (1) the state of speech recognition technology, (2) scoring the *Interactions* subskill, and (3) test difficulty. The goal is to address each of these topics and enable aviation professionals to make informed decisions about the role of automation in language testing.

The Nature of Automated Testing

When most people think of computerized testing, they envision an examinee in front of a computer. However, a computerized test need not involve a workstation, monitor or laptop; a computerized spoken test can be administered over the telephone. For example, to take a *Versant Aviation English Test*, an examinee dials a toll-free number and enters a unique Test Identification Number (TIN) using the telephone keypad. The system then plays instructions, which are printed verbatim on a test paper. The system presents items, one at a time, and records the examinee's responses. The system detects speech and can therefore determine when the examinee has stopped speaking and advance to the next item in the test. The responses are sent over a secure connection to a centralized database and scoring system. After scores are calculated, the score reports are posted to a secure web site. The entire process of scoring occurs within a few minutes after the examinee has completed the test.

The *Versant Aviation English Test* measures facility in spoken English in the aviation domain; that is, the ability to understand spoken English on work-related topics and to respond appropriately and intelligibly in spoken English at a fully-functional conversational pace.

Facility in spoken English is essential to successful radiotelephony communications – 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 communication, especially when attention is focused on aviation-related tasks at hand. Because the *Versant Aviation English Test* requires real-time language processing, it measures the degree of automaticity in the encoding and decoding of the core elements of oral language. Automaticity is the ability to access and retrieve lexical items, to build phrases and clause structures, and to articulate responses without conscious attention to the linguistic code (Cutler, 2003; Jescheniak, Hahne, and Schriefers, 2003; Levelt, 2001). Because cognitive capacity is limited, automaticity is required in order for the speaker/listener to be able to pay attention to the unfolding situation rather than paying attention to what needs to be said or how the encoded message should be structured.

Arguments Against Automation

In language proficiency testing, expert judgment has been considered the gold standard. With the introduction of automation, questions arise as to the suitability of employing a new technology as a replacement solution. Three areas of concern will be addressed: the accuracy of speech recognition technology, automation for the assessment of ICAO's *Interactions* subskill, and scoring severity.

Concern #1: Speech recognition technology does not work.

Automatic speech recognition technology has been deployed in many domains over the past few decades including dictation systems (e.g., ViaVoice™), call center applications (e.g., Amtrak, American

Airlines, UPS), and voice dialing on cell phones. Unfortunately, some users' experiences with these systems have been negative because of inconsistent recognition. Such experiences create a preconception of the limitations of speech recognition technology.

Over the past several years, automatic speech recognition has improved dramatically. Early systems were only able to handle simple yes/no responses. Now, the technology is capable of processing very complex sentence structures. For *Versant* tests, the acoustic models are different from those used in other speech recognition engines because they are trained specifically from samples of non-native English speakers. This means that the *Versant* tests are expecting (heavily) accented speech and can understand what non-native speakers are intending to say, regardless of poor pronunciation. The examinee's pronunciation is analyzed and scored separately.

In *Versant* tests, speech recognition, as such, is a small part of a larger system of spoken processing technologies integrated together to evaluate an examinee's spoken performance. Each response helps the system assign a level of speaking ability to the examinee. Since 54 items contribute to this process independently, the assigned score becomes very reliable as all the items are considered.

An empirical test of how well the speech recognition and speech processing technologies are performing is a comparison of scores generated from speech processing technologies with scores generated from human experts. Such an experiment was conducted with the *Versant for English* test. Test scores were gathered from examinees using the speech recognizer and other speech processing models. Separately, scores were generated for the same examinees from careful transcriptions and expert ratings of pronunciation and fluency. The correlation between the two sets of scores was 0.97. Such a high correlation indicates that machine scores are virtually indistinguishable from scoring that is done by careful human transcriptions and repeated independent human judgments. A scatter plot is shown in Figure 1.

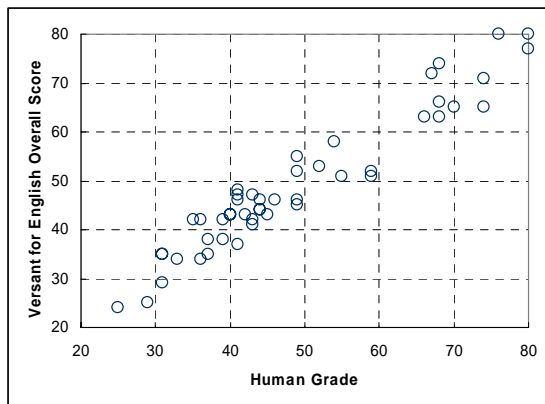


Figure 1. Scatter plot of machine generated Overall scores and human grades, $n=50$, $r=0.97$.

A similar correlation is expected for the *Versant Aviation English Test*, since both tests are built within the same technical framework.

Thus, the speech recognition technology used in *Versant* tests to assign a score of spoken language performance is extremely accurate as seen from comparisons with human-generated scores.

Concern #2: A computer cannot assess the subskill Interactions.

Another common concern is that ICAO's *Interactions* subskill cannot be measured by a machine. 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 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. But how would one know that one rater's perception of 'slow' is the same as another rater's? The *Versant* machine records precisely how many milliseconds it takes the test-taker to respond to an item. From the machine's point of view seven seconds is always seven seconds. Latency measurements are then compared to the same measurements from proficient speakers and then are rescaled. The calibrations are done the same way for each test taker, so there is no inconsistency or bias. In this way, the machine is more reliable than humans at

measuring response latency.

The assessment of whether the response is appropriate is simply a matter of identifying what the test taker has said and judging the appropriateness of the content of the response. This analysis can be done automatically by using speech processing technologies (recognizing what the test taker has said), and Item Response Theory (IRT)¹, which produces an estimate of the test taker's ability according to the item's difficulty. Here again, the analysis is based on more than just one person'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.

Other aspects of the *Interactions* rubric such as dealing with "an unexpected turn of events" and "checking, confirming, or clarifying" information are more closely associated with the tasks in the test as opposed to whether or not a machine can score the performance. The *Versant* test includes tasks that present predictable situations as well as unexpected events and provides ample opportunities for test-takers to deal with apparent misunderstandings by clarifying, confirming, and correcting information.

Aside from technical arguments, there is also a philosophical element to the issue. It is still difficult for many to profess that a machine might be as good if not better than a human being at performing complex tasks. The debate about whether or not machines can evaluate *Interactions*, which may involve not only skill but intuitive human judgment, is similar to the debate about whether a machine could ever beat a human player in chess. In 1995, the world chess champion, Garry Kasparov, insisted that no computer would ever beat him because, in his view, there is more to playing chess than mathematical calculations. He claimed that there is a certain 'art' to winning. Two years later, after a match against IBM's chess machine, Deep Blue, he

¹ IRT is a more modern approach to test theory, compared to classical test theory. In IRT, item parameters are related to examinee abilities and item responses.

was proven wrong.

The machine was not based on artificial intelligence and therefore did not approach the problem the same way a human would. Instead, the designers took advantage of what computers do best: apply massive computing power to a search problem. In a similar vein, an automated analysis of *Interactions* exploits the very things that machines do well. Computers are very good at measuring time, and automatic speech recognition can be characterized as a special type of search problem. In addition, scores can be scaled accurately because the computer encodes information about which items are surprising or difficult to non-natives, based on actual data from examinees.

The key dimensions that define the *Interactions* subskill are how immediate and appropriate the responses are. Because a machine can measure latency of speech and can evaluate whether or not a response is appropriate (in both predictable and unexpected situations), a machine can, in fact, measure ability on the *Interactions* scale.

Concern #3: The test will be too difficult for some non-native pilots and air traffic controllers.

In many countries, there is a concern that local pilots and air traffic controllers will not be able to achieve the proficiency levels specified by ICAO. To some, anxiety about this issue is heightened with the prospect of automatic scoring. To be clear, these two issues are independent of one another: the automatic nature of the test does not define the test's difficulty or the test's scale. It would be equally possible to create an automatically scored test that all examinees passed as it would be to create a test that all examinees failed. For the *Versant* test, each subskill is scaled separately, and the scaling is based on the ICAO rubrics. ICAO released rated speech samples to facilitate the implementation of the ICAO standards. These speech samples will help all test developers align themselves to the standards. Therefore, the issue for test publishers is not whether the scoring is too strict or too easy, but rather how well test scores align with the rated samples. The goal is to align with the ICAO Language Proficiency Standards as closely as possible.

Although there may be some hesitancy about using automatic scoring in language testing, many of the posed challenges are not causes for concern but rather a means for providing scoring advantages. For example, speech recognition technology used in *Versant* tests is very accurate because it can extract rich information about the speaker's responses in ways that had not been possible before. Some may say that *Interactions* cannot be scored by a machine, but the machine can actually produce more precise metrics of speed and appropriateness than subjective impressions. Finally, automatic scoring adds a level of consistency that can help the test align more reliably to the ICAO standards. In addition to these advantages, automation in language testing offers other significant benefits in accuracy, security, accountability, and efficiency.

Arguments in Favor of Automation

Automation provides many benefits for test administration and scoring, not only from the point of view of test reliability and validity, but also with regard to practicality.

For test administration, recorded voices are presented over a telephone. This administration paradigm enhances authenticity by offering a close simulation to radiotelephony communication. In both telephone and radiotelephony interactions, no facial cues or body language can aid communication. An important requirement in the ICAO standards is to measure communication in voice-only situations, which can be more challenging than face-to-face interactions (ICAO, 2004, Section 2.7.1).

Because the voices are all recorded, presentation of test material is consistent across test administrations. Nuances in intonation for each item are therefore preserved and reproduced across test administrations. The recordings also allow test items to be presented in different voices and therefore reflect a controlled range of accents and speaking styles of aviation professionals.

Automation allows test forms to be generated dynamically, which keeps items secure. In the *Versant Aviation English Test*, each unique test form is created

from a pseudo-random selection of items from a large item pool, limiting item exposure.

When considering practicality of test administration, the benefits of automation are clear. Most tests will require hiring, training and monitoring interviewers and raters. Finding qualified professionals will be a challenge, especially for organizations with large volumes of examinees. With automation, large-scale administrations are possible and ensure rating consistency.

Automation provides significant benefits not only for administration, but for scoring. The scoring models embody hundreds of judgments from expert raters who are familiar with both aviation and language testing. Therefore, examinees do not have to be concerned about the qualifications and training of their individual rater. Since all tests are scored by the same machine, scores are not affected by differences in rater severity or bias. Consistency and objectivity in machine scoring eliminate this source of measurement error.

Again, with respect to practicality, automation has its benefits. The automatic nature of the test allows examinees and aviation organizations to view scores immediately. This can help with efficient decision making and enhanced training.

Conclusion

This paper addresses three issues about automatically administered and scored English language tests in the aviation domain: the accuracy of speech recognition technology, the ability of a computer to score the subskill *Interactions*, and concerns of scoring strictness. In the context of *Versant* tests, speech recognition is almost indistinguishable from the best human scoring. With regard to the *Interactions* subskill, the rubrics emphasize response latency and the appropriateness of the response. Because a machine can measure latency of speech and evaluate whether a response is appropriate (in both predictable and unexpected situations), a machine can, in fact, score *Interactions* and can produce more precise metrics of speed and appropriateness than subjective impressions. With regard to scoring, the focus for the developers of the *Versant Aviation English*

Test is how well test scores align with the ICAO Language Proficiency Standards.

Automation in language testing offers many benefits. Telephone administration, for example, closely simulates radiotelephony communication; automatic administration obviates the need for large numbers of qualified interviewers and/or raters; and scores are not affected by differences in rater severity.

Understanding the technical details behind automated technology and its advantages can enable companies, organization and governments to make informed decisions about the role of automation in language testing. Regardless of whether or not an automated test is the right choice, any test that is selected should be reliable, valid and allow for practical implementation. If these criteria are met, then the international community will be one step closer to accomplishing ICAO's goal of improving safety through effective radiotelephony communications across the world.

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.
- Hopkins, M. (1995). Computer challenges world's chess champion. *Azerbaijan International*, 3(3).
- International Civil Aviation Organization (2004). Manual on the implementation of the ICAO language proficiency requirements, First Edition, Document 9835.
- 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. (2001). Spoken word production: A theory of lexical access. *PNAS*, 98(23), 13464-13471.
- Weber, B. (1997, May 12). IBM chess machine beats humanity's champ. *The New York Times*.