IMPLEMENTING ICAO LANGUAGE PROFICIENCY REQUIREMENTS IN THE VERSANT AVIATION ENGLISH TEST

***Alistair Van Moere**, Knowledge Technologies, Pearson*
Correspondence to Alistair Van Moere: avanmoere@pearson.com
***Masanori Suzuki**, Knowledge Technologies, Pearson*
Correspondence to Masanori Suzuki: masanori.suzuki@pearson.com
***Ryan Downey**, Knowledge Technologies, Pearson*
Correspondence to Ryan Downey: ryan.downey@pearson.com
***Jian Cheng**, Knowledge Technologies, Pearson*

Alistair Van Moere, Masanori Suzuki, Ryan Downey, and Jian Cheng work in research and test development at Knowledge Technologies, Pearson. They develop automatically-scored spoken and written language assessments for use globally in high-stakes contexts. Alistair Van Moere holds an MA from Warwick University and PhD from Lancaster University in Applied Linguistics. Masanori Suzuki holds an MA in TESOL from San Francisco State University. Ryan Downey holds a PhD in Language and Communicative Disorders from the University of California, San Diego and San Diego State University. Jian Cheng has a PhD in Artificial Intelligence from the University of Pittsburgh.

Correspondence to Jian Cheng: jian.cheng@pearson.com

This paper discusses the development of an assessment to satisfy the International Civil Aviation Organization (ICAO) Language Proficiency Requirements. The Versant Aviation English Test utilizes speech recognition technology and a computerized testing platform, such that test administration and scoring are fully automated. Developed in collaboration with the U.S. Federal Aviation Administration, this 25-minute test is delivered via a telephone or computer. Two issues of interest are discussed. The first concerns the practicalities of assessing candidates in each of six separate dimensions of spoken proficiency: Pronunciation, Structure, Vocabulary, Fluency, Comprehension, and Interactions. Although an automated scoring system can objectively segregate these skills, we question whether human raters have the capacity to do this in oral interviews. The second issue discussed is how an automated test can provide a valid assessment of spoken interactions. Tasks were designed to simulate the information exchange between pilots and controllers on which candidates' proficiency in 'Interactions' could be measured, for example, by eliciting functions such as correcting miscommunications and providing clarification. It is argued that candidate ability can be probed and estimated in a fair and standardized way by presenting a series of independent items which are targeted in difficulty at the various ICAO levels.

ARAL 32:3 (2009), 27.1-27.17. DOI 10.2104/aral0927
ISSN 0155–0640 / E-ISSN 1833–7139 © AUSTRALIAN REVIEW OF APPLIED LINGUISTICS

INTRODUCTION

This paper discusses issues in aviation English assessment with respect to the Versant Aviation English Test (VAET), which was developed under a co-operative research and development agreement with the Federal Aviation Administration (FAA) of the United States. The test is intended for use internationally to certify both pilots and air traffic controllers according to the resolution of the International Civil Aviation Organization (ICAO, 2004).

The constructs of oral performance tests are often defined in the rating criteria (Fulcher, 1996, p. 37), and indeed, the VAET was constructed based on ICAO's six-band, analytic language descriptors. However, although the ICAO Manual is informative in many areas, little information is provided about the development and theoretical rationale for the rating criteria. This can present challenges for test designers and for the aviation field in general, because poorly defined or poorly operationalised criteria are likely to result in different interpretations of the requirements. Different tests designed from the same criteria could exhibit variation with respect to task requirements, scoring, and standard-setting. Accordingly, this paper describes the development and validation approach that led to the VAET and discusses how two potentially challenging areas related to the ICAO language proficiency criteria were overcome.

The first area relates to ICAO's stipulation that a candidate's final ICAO level is equal to their level in the subskill in which they are least proficient. This stipulation requires that each of the six subskills must be disambiguated from the candidate's performance and measured analytically. Since pass-fail decisions could be made on the basis of ability on a single subskill, the test designer must demonstrate that each subskill is assessed reliably. The approach taken in the VAET to ensure separate measurement of subskills is by means of a computerised scoring algorithm which analyses a candidate's responses and feeds different elements of the speech into different subskill scores, thereby ensuring that there is little or no overlap in the scoring of subskills.

The second area discussed in this paper relates to the assessment of the interactions subskill. The candidate must be given an opportunity to demonstrate whether they can 'maintain exchanges even during an unexpected turn of events' (ICAO, 2004). Since the VAET is an automated test, the interactions subskill is assessed without the candidate participating in unfolding dialogue with an interlocutor. Rather, the candidate is assessed by eliciting independent performances on a series of routine and non-routine exchanges.

The paper is organised as follows. The next section describes the VAET tasks and test development. Following this is a discussion of the theoretical and practical challenges

associated with using the lowest of the six subskills to determine an ICAO level. We then address the assessment of the interactions subskill, and contrast the assessment of interactions in interview tests and in the automated test. This paper is not intended to provide a complete justification for making decisions on the basis of VAET scores, but rather to highlight several important challenges in aviation English assessment, inform stakeholders, and improve test standards.

VAET DESIGN AND DESCRIPTION

The VAET is a spoken English test that can be taken over the telephone or via computer, where test delivery and scoring are conducted by computer. During the test the system presents a series of recorded spoken prompts in English at a radiotelephony pace to elicit responses from the candidate. A recorded examiner's voice guides the candidate through the test, but instructions are also printed on the test paper or computer screen. Test items are presented in various native and non-native speaker voices that are distinct from the examiner voice. A test typically lasts 25 minutes and consists of eight sections with a total of 78 items. Scores for each subskill are reported on a scale of 10 to 70, where 10–19 is Level 1, 20–29 is Level 2, and so on. A description of the tasks is provided in Table 1 (for more information, see Pearson, 2008).

For candidates who answer all the items presented, the VAET extracts about six minutes of scored speech. This speech is dense with information about candidate ability; short three-second responses elicited in Repeats indicate whether the candidate has understood the prompt and how fluently and immediately they are able to formulate a grammatically correct response, while longer 30-second responses elicited in Story Retelling demonstrate the candidate's ability to control language in extended speech. The final task, Open Questions, is not scored at present but the candidate's responses are captured as audio files and made available for review by authorised test administrators.

Performance elicited in the VAET is scored using speech processing technologies. Briefly, the test development process was as follows. Tasks were designed to elicit the language, functions and cognitive demands of the target language use domain, such that the resulting performances could be scored with reference to the ICAO Language Proficiency Rating Scale. The test items were drafted and reviewed with reference to item specifications by experts with domain-specific knowledge. To ensure content relevance and representativeness, the language in the items was checked against an aviation corpus developed at Oklahoma State University. Acceptable item prompts were recorded by

aviation professionals in a recording studio. The prototype test was field-tested on 478 native speakers and 628 non-native speakers.

Task	Explanation
A. Aviation Reading	Read aloud sentences that are printed on the test paper or screen, one at a time, in the order requested.
B. Common English Reading	Prompt: 'Now read Sentence One'. Response: 'The flight was delayed because some technical issues needed to be resolved.'
C. Repeat	Listen to a sentence and repeat the sentence aloud word-for-word. Prompt: 'I usually get there a couple of hours before the scheduled departure time'. Response: 'I usually get there a couple of hours before the scheduled departure time.'
D. Short Answer Questions	Listen to a question and provide an answer with a word or phrase. Prompt: 'What is the name for land that is surrounded by water on three sides?' Response: 'A peninsula.'
E. Readback	Listen to a radiotelephony message and give an appropriate readback. Prompt: 'World Air 395, maintain flight level 070.' Response: 'Maintaining flight level 070, World Air 395.'
F. Corrections & Confirmations	Listen to a radiotelephony message and read the call sign on the test paper or screen. The message might contain correct or incorrect information, or a request for information. Respond appropriately. Prompt 1: 'World Air 395, continue descent to flight level 110, report passing 150.' Prompt 2: 'Descending to flight level 10 thousand, report passing 15 thousand, World Air 395.' Response: 'World Air 395, negative, continue descent to flight level 110, report passing 150.'
G. Story Retelling	Listen to an aviation scenario and then describe what happened in your own words. (Narratives are typically six to eight sentences long).
H. Open Questions	Listen to spoken questions and describe your own opinion or experience.

Table 1 Tasks, prompts and expected responses in the VAET

Field test candidates were all aviation professionals who represented a wide range of aviation experiences and 40 different first languages. Candidates' responses were captured as sound files and qualified raters subsequently evaluated the performance on each response and also gave overall holistic judgments of the candidates' ICAO level based on extended responses. Each response was rated by two or more raters independently. A total of 25,890 ratings were produced for the development of the automated scoring system and a further 20,673 were produced for the validation of the system. The items were analysed qualitatively for the language content elicited, and their psychometric properties were determined (difficulty, discrimination, precision) using Item Response Theory. Non-performing items were rejected, while high-performing items were assigned difficulty values which fed into the test scoring. Algorithms were created to predict expert ratings from characteristics in the candidates' spoken responses. For example, variables reflecting the pace, timing and pausing in the responses were combined to match the human ratings on the fluency criteria. Scoring models were then built which can process unseen candidate responses and simulate how they would be evaluated by expert human raters. A validation study using 140 unseen tests (i.e., tests not used for the scoring model development) confirmed that scores produced by the automated system were highly correlated with expert human judgments. Finally, the test was deployed onto a platform which enables test administrators to order tests, print out unique test papers, proctor the tests using telephone or computer, and then retrieve the candidates' test scores online minutes later. It is the administrator's responsibility to verify candidate identity and ensure that tests are conducted under exam conditions.

The next two sections each describe one test development challenge and show how they were overcome in the VAET. First, we address the analytic scoring of each of the six subskills.

SUBSKILL SEPARATION

ICAO stipulates that the candidate's final level is determined by their lowest performance among the six ICAO subskills, an approach that Bachman and Palmer (1996, p. 224) call *non-compensatory composite scoring*. This approach is ostensibly the most conservative and safest method to take when certifying candidates. However, from a test design and measurement perspective there are several practical challenges in its implementation.

One challenge is to establish whether these subskills actually exist as discrete, measurable traits within the language user in the first place. It would be incumbent upon the group setting the criteria to demonstrate that language users in the course of their real-

world tasks do not, in fact, achieve Operational level (i.e. ICAO level 4) or above by compensating across subskills. Take, for example, an aviation professional who is a ICAO level 4 or 5 in pronunciation, fluency, vocabulary and interactions, but with skills in structure at level 3. Would such a person be able to draw on their other skills in such a way as to compensate for their lack of grammar, such that a professional evaluator would, on balance, feel comfortable working with the candidate or certifying that person as level 4? Ideally, a source document such as ICAO's (2004) Manual – from which at least 20 different tests (Alderson, 2008) have been designed and developed in the last three years – should provide evidence or explanation of distinctness of traits.

Unfortunately, language testing researchers regularly fail to find the expected or posited distinctions between traits when those traits are quite similar, such as with dimensions of speaking ability (although it is possible to find clear distinctions between quite diverse traits such as reading and speaking, e.g., Bachman and Palmer, 1983). For example, using multi-trait multi-method analyses Henning (1987, p. 102) was unable to distinguish speaking traits such as fluency, pronunciation and grammar from test methods effects such as imitation, completion, and interview; the three traits correlated more highly with each other within test method than they did across test method. Hinofotis (1983) collected ratings on 12 performance traits and hypothesised three factors, but in fact found five factors, where pronunciation loaded onto a factor of its own, vocabulary and grammar loaded onto a single factor, and flow of speech (fluency) loaded together with traits such as confidence and presence.

Separation of subskills is particularly challenging when we consider the ICAO scales and the radiotelephony domain. A careful reading of the scales reveals an overlap in the operationalisation of subskills. For example, at level 4 the descriptors include:

- Comprehension, '(when confronted with) an unexpected turn of events, comprehension may be slower or require clarification strategies';
- Interactions, 'maintains exchanges even when dealing with an unexpected turn of events. Deals adequately with apparent misunderstanding by checking, confirming or clarifying';
- Vocabulary, 'can often paraphrase successfully when lacking vocabulary in unusual or unexpected circumstances'.

The functions for comprehending, clarifying and paraphrasing would appear to tap a similar underlying ability. This observation highlights the operational difficulties in

isolating six subskills of spoken language proficiency for the purpose of non-compensatory composite scoring.

Nevertheless, let us assume that even if there is overlap among subskills (and not adventitious covariance), each of the six contains elements of individuality such that they exist separately within the candidate and that candidates who are deficient in one subskill cannot safely be considered Operational. In this case the test designer should still address two other challenges. First, it should be ensured that the six subskills are treated equally with regard to their elicitation and the reliability of their scoring. This is because the test will only be as reliable as the weakest of the six measures that are assessed; if the subskill on which a candidate scores lowest is unreliable, the entire test is effectively unreliable. Second, it must be ensured that the six measures are indeed assessed separately. If raters transfer judgment across tasks by, for example, perceiving vocabulary as strong because of good grammar or by conflating pronunciation and fluency, then one of the subskills may not be properly assessed (McNamara, 1996, p. 216; Orr, 2002).

In an oral interview where the elicited performance is evaluated by a rater on all six subskills, reliability and discreteness may be established using two techniques:

1. by demonstrating that inter-rater reliability is equally high in each of the six subskills, and/or
2. by demonstrating that six different raters who each focus on only one subskill achieve comparably high inter-rater reliability with the ratings from a single rater who assigns ratings on all six subskills.

This challenge of non-compensatory composite scoring among subskills was addressed in the VAET by means of an *a priori* and *a posteriori* approach to development and validation (e.g., Weir, 2005). The *a priori* approach consisted of a theoretical rationale based on several principles to ensure subskills were measured separately. One principle was that expert human scoring of candidates' responses for the development of scoring models would occur discretely; human raters judged all responses for fluency, after which they judged performances for pronunciation, then vocabulary, and so on. This helped ensure that there was minimal transfer of judgment across subskills; raters only evaluated one trait at a time rather than several at a time, thereby lessening the chance that pronunciation and fluency ratings, for example, would be co-influential. A second principle was that in the scoring algorithms distinct information extracted by the speech processors from candidate responses would feed into different subskills. For example, if a type of

information from a response that was hypothesised to measure vocabulary fed into the vocabulary subskill, then it would not also feed into the grammar subskill.

The process of extracting information from responses to score subskills can be exemplified on the Readback task. The task feeds into the scoring of four of the subskills, but different information from the Readback responses is extracted for each subskill. For the scoring of the pronunciation subskill, the speech processors analyse the manner in which words in the response are articulated, including the lexical stress and segmental forms of the words spoken, as well as the pronunciations of the words within their phrasal context. For the fluency subskill, temporal information – including the rate of speaking, the position and length of pauses, rate and continuity of speech while using discourse markers – is extracted. For comprehension, the presence of certain words in the response demonstrates how well the candidate has understood the prompt. For interactions, the score is derived from the latency (immediacy) of the response as well as the completeness (informativeness) of the response. In this way, the candidate's performance is analysed and different pieces of information are partitioned out so that they feed into the scoring of the relevant subskills separately. In practice, absolute distinction cannot be achieved because some tasks naturally tap more than one subskill, but the theoretical principle was upheld as every subskill was informed by unique sources of response information. The scoring method which shows how task performance feeds into subskills is illustrated in Figure 1.

The *a posteriori* empirical validation consisted of two parts: first, an analysis to ensure that each subskill is scored reliably; second, an analysis of the dimensionality of the test scores via a correlation matrix. As mentioned in section 2, after the field testing, human rating and building of automatic scoring algorithms, tests from 140 candidates were used to validate whether the automated system assigned scores as expected. The 140 candidates were a diverse sample of the target population. The sample included candidates from 17 different language backgrounds who: were between 24 and 70 years old, split 80:20 male to female; occupied a variety of aviation professions; and had between one and 50 years of experience working in the aviation domain. The responses from these 140 tests had not been used in developing the recognition or scoring models. The tests were scored by the machine and also made available to expert human judges for transcribing and rating. Split-half reliabilities were calculated for both the automated scoring system and the human ratings. Table 2 presents the reliability coefficients for the subskills. They are consistently in the high range of r= 0.80 and above. To estimate the machine-to-human correlations, one of the human raters was selected randomly and their judgments were compared with the machine scores. Table 3 presents the intercorrelations

between pairs of subskills. As expected, the different dimensions of spoken language skills correlate highly but are sufficiently below unity to support the claim that the various subscores measure different dimensions of speaking ability.

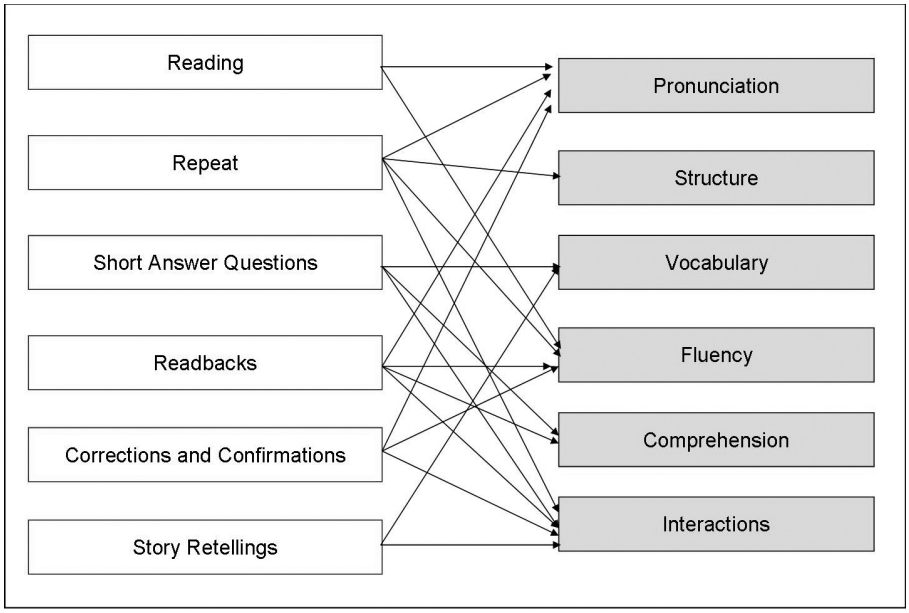

Figure 1 Relation of subscores to item types for the VAET.

The *a priori* theoretical rationale for development and *a posteriori* validation data reveal that the VAET is especially suited to fair and reliable non-compensatory composite scoring. The scoring algorithms are objective and dispassionate; a candidate whose performance falls below the pre-determined threshold will be scored accordingly, without reference to performance on the other subskills and without the bias or sympathies introduced by human judges.

The next section discusses the challenges associated with assessing ICAO's interactions subskill. The construct is operationalised with respect to the aviation domain, and the rationale for scoring this subskill in the VAET is described.

	Machine reliability	Human reliability	Machine–Human correlation
Pronunciation	0.97	0.98	0.89
Structure	0.82	0.84	0.93
Vocabulary	0.89	0.89	0.95
Fluency	0.95	0.97	0.82
Comprehension	0.88	0.87	0.94
Interactions	0.88	0.89	0.93
Overall	0.93	0.92	0.94

Table 2 Spearman-Brown split-half reliability coefficients and machine-to-human correlations coefficients for the VAET subskills and final score (adapted from Pearson, 2008)

	Pron.	Struc.	Vocab.	Fluency	Comp.	Interact.
Pronunciation	1					
Structure	0.72	1				
Vocabulary	0.75	0.70	1			
Fluency	0.90	0.69	0.67	1		
Comprehension	0.74	0.73	0.91	0.69	1	
Interactions	0.79	0.73	0.91	0.73	0.87	1
Overall	0.87	0.86	0.90	0.81	0.88	0.90

Table 3 Subskill correlations for automated scores (Pearson, 2008)

ASSESSMENT OF INTERACTIONS

For the test designer or evaluator, it is important to note that the operationalisation of interactions according to the ICAO criteria is quite different to *managing the conversational interaction* as defined elsewhere in the language testing literature (see below). The ICAO (2004) rating criteria define interactions largely in terms of 'immediate, appropriate, and informative' exchanges. In contrast, in everyday communication interactional competence is generally considered to consist of the ability to perform activities such as turn-taking, turn-allocating, topic introduction, ratification and topic-shifting (Schegloff et al., 1977). These activities often necessitate functions such as interrupting, requesting clarification, tactfully disagreeing, or rephrasing in order to check comprehension. But

researchers have cautioned that different speech events require different communication management strategies (Lantolf and Frawley, 1985). Even speech events in everyday communication exhibit extremes: information-related talk can be characterised by skills such as establishing common ground, giving information in bite-sized chunks, and clarifications, whereas speech events such as chatting can be characterised by the sustaining of social contact and may be strenuous at the social as well as linguistic level (Brown, Anderson, Shillcock, and Yule, 1984). These two kinds of interactions are so different that speakers may not be equally skilled at both kinds of talk.

The highly goal-oriented, formulaic functional interactions in radiotelephony (ICAO, 2004, Appendix B; Morrow, Rodvold, and Lee, 1994) are yet a different kind of speech event, and language test designers must put aside previous notions of 'competence in conversational interaction'. Interaction in the radiotelephony domain can be characterised as follows:

- *Reactivity* (responsiveness) and *goal-orientation* (purpose of the talk) (van Lier, 1989, p. 495) are very high; speech acts perform transactional roles that require a response.
- *Rights* and *duties* are equally distributed (Silverman, 1976, p. 142); both participants are responsible for efficient and successful communication, and for repair of breakdown.
- The conversational rules of turn-taking such as adjacency pairs, selection of the next speaker or self-selection, and competition for the floor (Schegloff et al., 1977) do not apply; turn-taking is formulaic and restricted by the medium of radiotelephony.
- Topics are not 'spontaneously created' or 'negotiated across turns', as in conversation (Sacks et al., 1974); rather, they are determined by unfolding events and information-transfer requirements.
- Politeness is less valued; direct utterances are more communicatively successful than mitigated utterances (Linde, 1988).
- Speakers do pick up elements of each others' contributions and incorporate them into their own speech (Brown and Yule, 1983, p. 89), but the purpose is to convey meaning efficiently and without ambiguity, rather than to show interest or friendliness (Tannen, 1984, p. 31).

Since the default method for assessing spoken interaction is commonly thought to be the interview, two approaches can be compared: human-administered interview, and machine-administered test items. Proponents of the interview method argue that it simulates natural, co-constructed interaction by participants (Swain, 2001). Interviews also

provide a somewhat realistic test-taking experience. Although communication in test settings is stilted and produced for the purpose of assessment rather than communication (Johnson, 2001), at least it occurs between real people and not between a person and machine.

But the interview approach also has disadvantages. In unscripted interviews the examiner can adapt to candidate requests and ask probing questions to establish the candidate's level, but might equally skew the assessment by accommodating to their level or taking too severe a line of questioning (Brown, 2003; Ross, 1992). If the interview or task is scripted in order to standardise the assessment, then co-construction is in any case illusory: the candidate would still be expected to provide a particular answer in response to a pre-determined question, and even to use pre-determined formulaic expressions. The candidate may request repetition or clarification, but this is relatively straightforward in radiotelephony (e.g., the candidate gives their call sign, followed by 'say again') and does not distinguish skill at interactions at anything above ICAO level 2. Further, a request for clarification might be used for any number of reasons: if it is to buy time for the candidate in order for them to formulate their response, then it should be reflected in their fluency score; if it is because the candidate did not understand the input, then it should be reflected in their comprehension score; if it is because the candidate did not understand one particular word in the statement, then it should be reflected in their vocabulary score. In sum, having an examiner or interlocutor who responds intelligently to the candidate may facilitate a satisfying test-taking experience, but it does not necessarily facilitate the measurement of interactional competence, and is most likely to create unequal test conditions across test occasions and test examiners.

Regarding automated test administration, interaction is assessed by presenting a series of independent items designed to elicit the desired linguistic and cognitive functions. The candidate's performance across all the items is then evaluated in terms of success with respect to the ICAO interactions descriptors. The advantage to this approach is that the candidate is given numerous opportunities to demonstrate their interactional competence. In the Readback task and Corrections & Confirmations task, each candidate is given a total of 20 opportunities to display interactional competence in response to a variety of prompts. The items have established difficulty values based on field testing and IRT analysis; the candidate can still score highly even if they fail to perform properly on one or more items. By presenting easy and hard items, it is ensured that the candidate's ability is properly probed and that each candidate is presented with enough items at their level to accurately estimate their ability.

In terms of disadvantages, the automated administration does not present the candidate with reciprocation or follow-up from an interlocutor. From a measurement perspective, however, this is less important; it is the immediacy, appropriacy and informativeness of the candidate's response that is being assessed, and not the interlocutor's rejoinder. It may also be the case that on balance, many candidates would prefer a human-administered interview (Qian, 2009), perhaps because they perceive it as easier or because they prefer the reassurance of an acknowledgment of their utterances. Accordingly, test-taker preferences are one factor that should be taken into consideration when selecting a valid test, along with other factors such as practicality, standardisation of administration, and reliable scoring.

INTERACTIONS IN THE VAET

The assessment of interactions in the VAET also followed an *a priori* and *a posteriori* approach. The *a priori* rationale for the tasks followed Weir's framework (Weir, 2005, p. 222). That is, the tasks were designed with the specific purpose of eliciting the construct and functions given in the descriptors; they were based on linguistic theory and evidence (Pearson, 2008, p. 10); they were evaluated by aviation experts from the FAA and other international aviation experts; and they contained language content that was informed with reference to a corpus including authentic radiotelephony speech. Two tasks, Readbacks and Corrections & Confirmations, were specifically developed to inform the scoring of the interactions subskill.

A careful reading of the ICAO interactions descriptors reveals that the subskill can be reduced to several elements. The elements that emerge from levels 2 to 5 are 'immediate, appropriate, and informative' responses, and 'initiating and maintaining exchanges'. Further, the level 4 descriptors include three key interactional functions by which these elements might be accomplished: 'checking, confirming and clarifying'. As with all the subskills, the distinguishing characteristic between level 3 and level 4 is that a level 3 performance is defined by proficiency in routine or predictable situations, whereas a level 4 performance is additionally defined by proficiency in unusual or unexpected circumstances. In other words, interactions is the ability to initiate, maintain and respond, and to do so immediately, appropriately, and informatively using both phraseologies and common English.

Briefly, these elements are extracted from candidates' responses in the VAET as follows (cf. Pearson, 2008). Immediacy was operationalised as the latency, or delay, in the candidate's responses. Illustrating the importance of modeling inappropriate silence, the

ICAO manual (2004, Appendix A:14) provides one sentence to describe the Interactions subskill: 'Pilots and controllers should be aware that inappropriate silence may indicate a failure to understand.' Immediacy of responses is measured from Repeat, Readback, and Corrections & Confirmations tasks, thereby ensuring that the response latency is extracted from various contexts using routine and non-routine language.

Appropriateness and informativeness of responses were operationalised based on the content of the responses: whether certain keywords were present and the completeness of the response. Since the interactions subskill must consist of ICAO phraseology and common English in routine and non-routine contexts, the scoring algorithms for interactions include information not only from Readback and Corrections & Confirmations, but also from Repeats, Short Answer Questions, and Story-Retelling.

The *a posteriori* validation consisted of linguistic and statistical analysis of the field tests. The linguistic analysis involved transcribing and analysing a large number of native-speaker responses to verify that native speakers were able to respond as expected. If less than 80% of native speakers were able to answer correctly, the items were discarded. In some cases, native speakers provided unexpected responses which were judged by experts as correct, and these responses were added to the answer keys. Statistical analysis involved IRT modeling to verify that items: 1) met an expected distribution of difficulty with regards to the target population; 2) exhibited unidimensionality in that they all measured the same ability in the candidates; and 3) discriminated among candidates according to the ability to be measured.

Tables 4 and 5 show the psychometric properties of the Readback and Corrections & Confirmations items based on the field testing; these items currently reside in the VAET question pools. The measure column indicates the range of item difficulty in logits and reveals that items were of varied difficulty, and therefore suitable for accurately evaluating a range of candidates according to ICAO levels 1 - 6. The Infit Mean Square column shows that items fell between the values of 0.6 and 1.4, and therefore tapped the same ability in candidates. The Point-biserial column indicates that all items have values above 0.3 and are therefore powerful at discriminating among candidates according to ability. Scaling was achieved just as for the other subskills: by collecting human expert assessments of candidate performance and mapping to candidate ability on the ICAO scales.

	Measure	Infit MnSq	Ptbis
Mean	1.11	0.92	0.49
Standard Deviation	0.55	0.16	0.11
Minimum	-0.04	0.62	0.30
Maximum	2.80	1.35	0.76

Table 4 Item properties for Readbacks

	Measure	Infit MnSq	Ptbis
Mean	0.75	0.98	0.43
Standard Deviation	0.53	0.09	0.09
Minimum	-0.76	0.78	0.30
Maximum	1.37	1.21	0.58

Table 5 Item properties for Corrections & Confirmations

CONCLUSION

This paper has addressed three main topics. It has provided an overview of the test development process for the Versant Aviation English Test. It has explained how candidate performances are analytically processed following theoretical principles, and how features of spoken responses are partitioned for scoring the six subskills. It has also explained how the performances elicited in certain tasks are representative of the interactions construct.

The paper has also compared an automated and human-administered interview test. It has been posited that, when assessing six subskills separately in a fair and standardised way the interview approach is fraught with difficulties which, while not insurmountable, may require careful validation and then an ongoing standardisation of scoring. Machine scoring, on the other hand, is suited to multi-skill, non-compensatory scoring as it distinguishes performance data according to theoretical principles and evaluates it based on empirical data from thousands of field test responses.

We acknowledge that one limitation to automated assessment is that the candidate does not engage in the co-construction of naturally evolving communication. However, this is countered by the observation that such co-constructions in interviews are actually staged and that the candidate can only answer in a restricted number of ways, especially

in radiotelephony. We further argue that numerous independent, trialed and standardised items, which are presented in various speech styles and which elicit the intended interactions, are at least as fair and reliable for assessing interactional competence as compared to a series of questions posed by local examiners who must then multi-task and make an evaluative judgment on the candidate's interactional competence as distinguished from the other five dimensions of spoken ability.

Looking forward, we see several beneficial areas for research in aviation English testing. This paper has already pointed toward the need for a validation of the existing language proficiency scales. We also see a need for a standard-setting exercise which produces benchmark samples of the ICAO levels, and which involves many of the available language tests so that it can be verified that standards are applied equally rigorously, and that standards do not vary over time. Such a research agenda would build upon the foundation already laid by ICAO, and would allow for further improvements in aviation English assessment and air traffic safety.

REFERENCES

Alderson, J. Charles (2008). *Final report on a survey of aviation English tests*. Unpublished report.

Bachman, Lyle F; Palmer, Adrian S (1983). The construct validity of the FSI oral interview. *Language Learning*, 31(1), 67–86.

Bachman, Lyle F.; Palmer, Adrian S. (1996). *Language Testing in Practice*. Oxford. Oxford University Press.

Brown, Annie (2003). Interviewer variation and the co–construction of speaking proficiency. *Language Testing*, 20(1), 1–25.

Brown, Gillian; Yule, George (1983). *Discourse Analysis*. Cambridge: Cambridge University Press.

Brown, Gillian; Anderson, Anne; Shillcock, Richard; Yule, George (1984). *Teaching Talk: Strategies for Production and Assessment*. Cambridge: Cambridge University Press.

Fulcher, Glenn (1996). Testing tasks: Issues in task design and the group oral. *Language Testing*, 13(1), 23–51.

Henning, Grant (1987). *A Guide to Language Testing*. Boston, MA: Heinle and Heinle.

Hinofotis, Frances B. (1983). The structure of oral communication is an educational environment: a comparison of factor–analytic rotational procedures. In Oller, J.W., Jr (Ed.), *Issues in Language Testing Research* (pp. 170–187). Rowley, MA: Newbury House.

International Civil Aviation Organization (ICAO) (2004). *Manual on the Implementing of ICAO Language Proficiency Requirements*. Montreal: ICAO.

Johnson, Marysia (2001). *The Art of Non–conversation: A Reexamination of the Validity of the Oral Proficiency Interview*. New Haven, CT: Yale University Press.

Lantolf, James; Frawley, William (1985). Oral proficiency testing: A critical analysis. *Modern Language Journal*, 69(4), 337–345.

Linde, Charlotte (1988). The quantitative study of communicative success: Politeness and accidents in aviation discourse. *Language in Society*, 17(3), 375–399

McNamara, Tim (1996). *Measuring Second Language Performance*. London: Longman.

Morrow, Daniel G; Rodvold, Michelle; Lee, Alfred (1994). Nonroutine transactions in controller–pilot communication. *Discourse Processes*, 17(2), 235–258.

Orr, Michael (2002). The FCE speaking test: Using rater reports to help interpret test scores. *System*, 30(2),143–54.

Pearson, 2008. *Versant Aviation English Test: Test description and validation summary*. Palo Alto, CA: Author.

Qian, David (2009). Comparing direct and semi-direct modes for speaking assessment: Affective effects on test-takers. *Language Assessment Quarterly*, 6(2), 113–125.

Ross, Steven (1992). Accommodative questions in oral proficiency interviews. *Language Testing*, 9(2), 173–186.

Sacks, Harvey; Schegloff, Emanuel; Jefferson, Gail (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4), 696–735.

Schegloff, Emanuel; Jefferson, Gail; Sacks, Harvey (1977). The preference for self-correction in the organization of repair in conversation. *Language*, 53(2), 361–382.

Silverman, David (1976). Interview talk: Bringing off a research instrument. In Silverman, D. and Jones, J. (Eds.), *Organisational Work: The language of grading, the grading of language* (pp. 133–150). London: Collier Macmillan.

Swain, Merrill (2001). Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores. *Language Testing*, 18(3), 275–302.

Tannen, Deborah (1984). *Conversational Style: Analyzing Talk among Friends*. Norwood, NJ: Ablex.

van Lier, L (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly*, 23(3), 489–508.

Weir, Cyril J. (2005). *Language Testing and Validation: An Evidence-based Approach*. New York: Palgrave Macmillan.

Cite this article as: Alistair Van Moere et al. (2009). 'Implementing ICAO Language Proficiency Requirements in the Versant Aviation English Test'. *Australian Review of Applied Linguistics* 32 (3), 27.1–27.17. DOI: 10.2104/aral0927.