

Proceedings of the 11th International Civil Aviation English Association Forum

Rating for ICAO Language Proficiency Standards

Hosted by the Ecole Nationale de l'Aviation Civile, Toulouse

6th and 7th September 2011

The association was fortunate enough to be able to inaugurate ENAC's brand new auditorium for the plenary sessions which were therefore held in optimum conditions of comfort and ease of communication. Classrooms were used for the breakout sessions. ENAC teachers facilitated the question and answer sessions and movement between the auditorium, the breakout rooms and the restaurant.

Tuesday 6th September

We were welcomed by **Pascal Revel**, Vice-President of the ENAC and **Michael O'Donoghue**, head of ENAC's English department.

Moderator: **Philip Shawcross**

This Forum was an opportunity to mark ICAEA's 20th anniversary and the ICAEA President reminded delegates just how appropriate a venue ENAC was with its long tradition of aviation English starting with the pioneering and seminal works of **Yves Rengade**, **Jeremy Mell**'s doctoral thesis on radiotelephony communications (*Etude des communications verbales entre pilote et contrôleur en situation standard et non-standard*) and his leading role within the PRICESG (Proficiency Requirements in Common English Study Group). This tradition of participating in the development of aviation English at ENAC has been continued with **Mike O'Donoghue**'s involvement in the first version of the ICAO Rated Speech Sample Training Aid CD and the development of a new proficiency test, and **John Kennedy**'s teacher training expertise. It was also an opportunity to remember that it was 27 years ago that **Fiona Robertson** and her colleague **Joan Bellec** of the CLA (Centre de Linguistique Appliquée) of the University of Franche Comté, France, had organised the first gathering of aviation English teachers in Paris Orly; the association is very grateful to her for having practised her organisational skills again in preparing this 11th Forum with the assistance of the ENAC team and **Bozena Slawinska**.

The delegates were also warmly thanked for their presence and reminded that the registration fees they paid for attending ICAEA events were virtually the association's only source of income and that it was this income which allowed association to fund its activities: starting the ICAO Rated Speech Sample project, taking part in the ICAO test endorsement working group, creating the on-line ICAEA Information Resource, running a website, organising events, supporting LPR implementation, drafting ICAO circular 323, cooperating with other associations (ILTA, ALTE, EAMTC etc.) and covering its administrative costs. All the time devoted by the Bureau and Board members was on a purely voluntary basis. This income would also allow the association to pursue its plans for creating a virtual arena for university research and providing teacher and rater training and aviation English expertise in the developing world.

It was pointed out that the purpose of the Forum was to explore two issues which were pivotal for Language Proficiency Requirement implementation: rating standardisation and test endorsement. These were also two examples of close cooperation between ICAO and ICAEA, a cooperation which had started with the 8th Forum in Warsaw in 2002.

We regretted the absence of **Henry Emery** who as Rated Speech Sample project leader had devoted so much time and energy to making the training aid a reality and that of

Professor Charles Alderson and **Dr. Andrea Revesz** of Lancaster University, and **Dr. Ute Knoch** of the University of Melbourne, whose contributions had been invaluable. ICAEA was extremely grateful to Henry, Charles, Andrea and Ute and Cambridge ESOL for their support.

Michael O'Donoghue, ENAC: *Review of ICAO LPR testing terminology*.

Michael usefully set the framework for the forum by retracing the history of aviation English from Yves Rengade's joining the ENAC in 1968, through the work of the PRICESG and the ICAO Language Proficiency Requirement standards and recommended practices to the current measures to enhance implementation. He went on to characterise the distinction between phraseology and plain language referring to Doc. 9835 3.3.14 and Circular 323 4.2.9. He reminded us that proficiency tests should be based on effective communication and dialogue management, not answering discrete item questions. The jury was still out on broad and narrow interpretations of the aviation content of test tasks. He indicated that there was always an element of subjectivity in rating and this comment was later echoed by other speakers.

Angela French, Dr. Evelina Galaczi, Cambridge ESOL, **Mike French**: *The ICAO Speech Samples Rater Training Project*.

The purpose of the project, the members of the panel, the time line of the process, the number of raters (75), the number of countries involved (23) and the number of samples (210) were presented.

Mike French presented the development of the web interface and database; initially there were many unknowns, such as the number and size of the samples. The sound quality of the samples provided varied greatly and required considerable cleaning up and editing. For reasons of security and expediency, it was decided to assign a dedicated area to each rater rather than have a central database. This ensured autonomy of operation, with no interaction between raters.

Mike emphasised the good nature and dedication of all the raters involved in the project, and the amount of e-mail communication that ensued has provided significant insight into the problems faced by raters in the course of their work.

Evelina gave a masterly explanation of rater tendencies (harshness and leniency) and interpretation (central tendency and halo effect) before going on to explain FACETS analysis, the data which could be obtained from it and its application to the ICAO speech samples project. This data was extremely useful in monitoring raters both individually and as a group.

The first phase of the training aid containing 22 samples has been delivered to ICAO and will soon be downloadable free of charge from the ICAO website. A second batch of samples is currently being processed and will be uploaded to consolidate the training aid as soon as it is complete.

Tuesday breakout sessions

Michael Kay, Australian Council for Educational Research: *Initial and recurrent rater training*.

The purpose of training was to calibrate raters and reduce errors. Michael recommended at least 5 continuous days of training in a group of not more than 12. It was important not to focus on negatives, but rather ask what this candidate can do. During the training it was also important for the raters to play the role of candidate in mock tests for a better

understanding of the process and the test-taker's experience. 60% of the course should be spent rating samples. Borderline cases were particularly instructive. Raters should be encouraged to change their mind and not stick to first impressions if the candidate's performance varies during the test.

A question was raised about rating comprehension: should it be done independently or through interaction? It was suggested that above Level 4 a range of different voices was required.

John Kennedy, ENAC: *Inter-rater reliability*.

John presented the process used at ENAC to develop the MTF test which was rated remotely and independently. Candidates appreciated the fact that the interlocutor was not a rater. The disadvantage of remote rating was a two-week turnaround and a lot of file administration. The first raters tended to be a little more conservative about awarding Level 5 than the third rater who was called in if there was disagreement. It was found that language experts and operational raters tended to rate in a similar way. Inter-rater reliability was in the region of 95% for candidates in the middle of a band.

The test was mainly designed for ENAC students. It had taken 15 months to develop and was used for Levels 4 and 5 only. There was a 30-minute sound file for each candidate of which 22 minutes were of the candidate speaking.

Sergey Melnichenko, Complang Training Centre: *Another look at testing and rating validation*.

Sergey reminded us that the integrity of the test, of the testing process and of the rating had to be ensured. A test cannot be valid but unreliable, however it can be reliable but not valid, i.e. not measuring what it should.

Although pilots and controllers share communicative macro-functions (see ICAO Doc. 9835 Appendix B and Doc. 4444 12.2.2), they also have specific micro-functions which should be represented in their dedicated tests. However, the main emphasis is on the shared macro-functions.

If one refers to the ICAO rating scale band descriptors, a test based purely on R/T can only allow up to Level 4 to be awarded; in a test based purely on phraseology, only Level 2 could be awarded.

Other points which Sergey made included: in listening comprehension, speakers at Level 4 should be chosen for inputs; there has to be a sufficient length of ratable language; a distinction should be made between an interactional and a transactional use of language; communicative skills should be evaluated using communicative tasks; retired SMEs (Subject Matter Experts) such as pilots or controllers may not be reliable if used for their expertise, as operational practice changes rapidly.

Sergey ended by reminding us that 70% of pilots and ATCOs do not comply with standard phraseology.

Angela French, Dr. Evelina Galaczi, Cambridge ESOL, **Mike French**: *Rating exercises using RSS training aid samples and others*.

Rating sheets were distributed to the delegates in the amphitheatre and each delegate was assigned an anonymous ID. Then two warm-up recordings were played as a brief calibration session. (Angela pointed out that normally a much larger number of practice samples together with their official marks would be played and discussed before beginning rater certification.) Following this familiarisation process, seven short speech samples were

played and rated individually by each delegate.

The team showed great devotion to duty and processed all the ratings overnight and performed a FACETS analysis which was presented to the delegates the next day.

The first day's proceeding closed at 6 p.m.

Tuesday evening

The ENAC organised a superb Garonne River and canal boat ride for all the delegates and **Yves Rengade** as an honoured guest; Yves has just published a new book with Cépadues, *Manuel d'anglais du pilote de ligne*. There was a generous supply of delicious canapés and refreshments. Toulouse laid on a warm evening and beautiful sunlight.

After the boat ride many of the delegates met for dinner in a bistrot on the banks of the Garonne.

Wednesday 7th September

Three breakout sessions to rate different test samples were conducted by **John Kennedy, Sergey Melnichenko and Michael Kay**

This was followed by a *question and answer session* in the amphitheatre

Q: When would there be a single universal test?

Sergey: ICAO has never had the finances or the human resources to produce a test, nor is it in their remit. Since 2003, the market place has been filled with tests of varying quality. Today it would be unacceptable for one test to eliminate all others.

Q: What was the workshop facilitators' opinion of the ratings given by the delegates?

Michael: There is never a single correct answer. It should also be remembered that sample ratings do not reflect what a complete rating would have been. The rating of pronunciation was particularly subjective as it depended so much on the rater's familiarity with certain accents. The marking of pronunciation often contaminated the marking of other criteria. It was important to always refer back to the scale.

Sergey: In the last analysis the acid test was: do you trust this pilot; would you fly with him/her? The intention of the test task in the sample played by Sergey was to put the candidate under pressure.

One delegate found the questions irrelevant and misleading and had rated the candidate Level 4; he would have done better in R/T.

It was pointed out by several participants that rating cannot be based on supposition, but only on evidence. It was up to the candidate to give the assessor a sample of language; no assumptions could be made.

Another delegate felt that candidates should have only operational staff, not linguists, as interlocutors.

Angela French, Dr. Evelina Galaczi, Cambridge ESOL: *Feedback from analysis of the previous day's rating.*

The team pointed out that the conditions were not ideal, but that the exercise was simply designed to show what could be done, what data could be extracted and what data could

be given to raters. Three charts were presented displaying data from the 75 delegate raters.

'+' indicated harsh and '-' lenient. A range of +1.5 to -1.5 was acceptable. Raters falling outside these parameters might benefit from extra training.

Consistency was the most important variable with infit values of 0.6 to 1.4 being an acceptable range. Below 0.6 demonstrated too little variability in rating. With a low infit (0.6 to 0.4), raters were possibly giving flat profiles or 'playing it safe' and awarding ratings from the middle of the scale only. At the other end of the scale, a value higher than 2 was considered too erratic and characteristic of someone who was unfamiliar with the scale.

The team reminded us that you cannot just rely on numbers; there is not always a right answer or rating and it's important to consider statistical results and expert judgement in a complementary fashion.

Jean-Marc Guillemette, ICAO project manager: *Test endorsement.*

Jean-Marc presented the purpose of the test endorsement process: "systematically reviewing and confirming that a test complies with ICAO LPRs"

The various pages of the website were presented and described and samples of the on-line questionnaire given.

The initial endorsement team comprised representatives from IFALPA, IFATCA, ILTA and ICAEA. After discussions which took place at the ICAO Paris workshop in December 2010, the team was opened up to include specialists with testing experience from Europe, Russia, China, Brazil and Argentina.

Endorsement requires the test service provider to make a large amount of documentation, including sample test recordings and rating sheets, available for the evaluators.

The team has just completed its assessment of the first "pilot" (trial) test and is engaged on a second. This will enable the endorsement process to be validated and enhanced. The process will be ready and open to the public on 3rd October 2011.

A discussion followed on the logistics and pricing of the endorsement process. **Jean-Marc Guillemette, Angela French, Mike French, Sergey Melnichenko, Michael Kay, John Kennedy and Philip Shawcross** : *Q & A round table*

The following topics were raised by the delegates:

- There was general concern about the number of bogus attributions of Level 6 in many countries given that under current legislation there is no requirement for retesting at Level 6. A new retesting requirement with every 2 years at Level 4, every 4 years at Level 5 and every 6 years at Level 6 was suggested.
- Interest was expressed in pursuing research into the characteristics of Level 6, their definition on the Rating Scale and how Level 6 can best be assessed
- There was considerable discussion about the justification for a \$3 unit endorsement fee for each test delivered in addition to the one-off \$5,000 fee. Jean-Marc explained that both the initial investment costs and the ongoing administration costs had to be recovered and that the endorsement operation needed to be cost-neutral for ICAO. The case of tests developed by CAAs and State organisations was raised.
- One delegate wondered whether there would be ongoing monitoring of endorsed tests. Jean-Marc replied that tests would undergo a shorter re-endorsement process every

three years, but that in the case of suspicious feedback an investigation could be triggered.

- There were questions about the washback effect of test endorsement. It was generally felt that these could not be predicted with any certainty, but that in all probability the endorsement process would drive up standards of test design and administration through peer pressure (i.e. the desire to have an endorsed test or an authority requiring any of their accredited tests to be endorsed) and would also raise awareness among decision-makers about the criteria which a valid and reliable test should demonstrate. From its inception the endorsement process had been conceived of as a formative process.
- It was pointed out that the eight years since the publication of the LPRs in March 2003 had been a steep and positive learning curve not only on the part of teachers and test service providers, but also on the part of regulators and managers. Aviation English was starting to enter a more mature phase in which there was a more widespread understanding of good practice.
- There was a call for an organ for more effective communication among ICAEA members. ICAEA undertook to create such a site for discussion with dedicated topic streams as required, such as LPR implementation, rating, teacher training, English in aircraft maintenance, test endorsement, LPR washback, the communication needs of airport staff, human factors etc. In the meantime an ICAEA Facebook site existed and was presented by **Gabor Sipos**. **Jean-Marc Guillemette** suggested investigating communities of practice.

The Forum came to an end at 5 p.m. with very supportive and enthusiastic closing remarks by **Marc Houalla**, the Director of ENAC.

ICAEA would like to express its gratitude again to ENAC for the exemplary way in which this forum was hosted with care, warmth and generosity.