

**Minutes of the first meeting of the Rated Speech Sample (RSS) project panel held in the Cambridge Assessment building, Park Terrace, Cambridge and hosted by Cambridge ESOL on Friday 30<sup>th</sup> October 2009.**

The meeting was attended by:

Professor Charles Alderson (CA), Nicole Barrette-Sabourin (NB), Captain Sutas Dejkunjorn (SD), Henry Emery (HE), Angela French (Af), Terence Gerighty (TG), Philip Shawcross (PS), Gabor Sipos (GS), Bozena Slawinska (BS) and, by conference call, Ronald Vega Bolanos (RVB).

Captain Herbert Schwarz (HS) was unable to attend at the last moment.

Ute Knoch (UK), Captain Michel Menestrot (MM) and René Sanchez (RS) are also members of the panel who did not attend this meeting.

The meeting opened at 9 a.m., was chaired by Henry Emery and reported by Philip Shawcross.

HE opened the meeting by thanking Cambridge ESOL for their hospitality, recalling how there was a need for a tool to enhance rating standardisation and how ICAO's involvement in the presence of NB was important. The agenda was presented:

## **1. Speech samples**

- 1.1 Type of test
- 1.2 Test quality
- 1.3 Which tests?
- 1.4 Test materials and documents
- 1.5 Criteria and levels
- 1.6 Quantity of samples
- 1.7 Tasks and samples
- 1.8 Length of samples
- 1.9 Testing comprehension
- 1.10 Interlocutor behaviour
- 1.11 Radiotelephony phraseology content
- 1.12 Types of task

## **2. Raters**

- 2.1 Criteria for rater selection
- 2.2 Determining language proficiency
- 2.3 Rater guidance
- 2.4 Data collection

## **3. Project specifications**

-----

The panel was provided with the following working documents:

- Annotated Rated Speech Samples Project Discussion Paper 1
- Extract from ICAO Circular 318 (6.4 Rater team expertise)
- Types of task checklist
- ALTE Individual Component Speaking Checklist
- The CEFR Grid for speaking developed by ALTE members
- Guidelines for interlocutor behaviour for examinations conducted in the individual mode
- Extracts from 'Into Europe – The Speaking Handbook' pp. 162-163: How to use the rating scale
- The Model Aviation English Test Checklist
- ICAO European Air Navigation Planning Group EANPG48 Appendix J 2006

The aims of the meeting were recalled, i.e. to define:

1. The criteria for selection of speech samples
2. Rater qualifications
3. Data gathering
4. The sample and data gathering mechanism
5. The supporting documentation for test providers and raters

and also to review and amend the discussion paper and the specifications document. Some questions in the discussion paper would be impossible to answer, others would answer themselves in the course of the meeting and in some cases there was insufficient information to answer the questions yet.

### **1.1 Type of test**

It was agreed that only speech samples taken from licensing tests should be used.

### **1.2 Test quality**

There was some discussion about the fact that selecting samples from specific tests would be perceived as an indirect approval although this was not the intention and there would be an explicit disclaimer to this effect. This issue would require careful handling.

It was wondered whether demonstration of validity should be requested. As there were relatively few providers who had published proof of validity, it was acknowledged that this would lead to the exclusion of some widely recognized tests and constituted in fact a separate issue.

CA pointed out that evidence of validity could be gathered separately and that the panel's efforts should be focused on the assessment of tasks not tests. It would be wise not to ask for proof of validity, but that the appropriateness of tasks had to be screened.

NB said that there was a need for a list of test providers.

Af felt that the panel should seek a wide range of test providers in order to procure a wide range of national samples.

GS suggested the use of pre-prepared questions as he wondered whether test providers would be willing to give real samples, but Af was of the opinion that using real material would be more beneficial for the community.

There was a decision to use actual tests rather than generating a new test.

BS reminded the panel that there were many in-house tests which could be appropriate.

There ensued a discussion about the appropriate length of speech samples for the training aid. Af pointed out that a sample could be short or long provided it generated the appropriate language. It was agreed that the panel could not prescribe the length, but that complete tests should be requested and that extracts or tasks would be selected from the test. It would be necessary to request permission to select an extract or task from a complete test sample.

NB observed that the use of the sample for training purposes should be disassociated from the actual rating of an individual for licensing purposes. CA added that the panel would be showing performances in a given task, not demonstrating a test-taker's performance in a test as a whole. Any one task could not demonstrate all language skill criteria. This was in fact an advantage because it would focus the training aid's users on particular skills.

Af pointed out the need for a checklist which enabled questions like 'Does this task demonstrate a, b, c, d, e etc?' when the samples were screened.

BS reminded the panel that people from an operational background might want to see the whole performance and find chunked extracts focusing on specific skills too analytical. This might affect the training aid's face validity for operational staff.

HE pointed out that it would be too long to develop a screening checklist during the meeting, but that it would be done internally.

### **1.3 Which tests?**

TG expressed the view that it was important to be as inclusive as possible to encourage a high input and HE said the priority was to be reassuring for test providers.

GS reminded the panel that (in Europe) the speech samples belonged to the individual CAAs. There were Eurocontrol training samples generated by workshop attendees in Luxembourg at level 5. In the case of ELPAC, it would be necessary to: 1) ask authorisation from Eurocontrol, 2) approach individual State users, 3) use the on-line sample versions with new test-takers. It would be problematic and costly.

TG spoke about RELTA although he could not represent them in any way; he recalled that there were 4 on-line sample Tests. He thought it would not be an option if more work and expenses were involved in providing samples although Michael Kay had expressed his willingness to help and there existed more than 14,000 samples on the Melbourne database.

NB confirmed that in theory all speech samples belonged to the national authorities. It would be necessary to approach the focus points at the CAAs mentioned on the ICAO fsix website.

CA did not believe that the test-takers' consent was required to use the samples they appeared in, but there was not agreement on this point.

HE raised the question of who actually needed this training aid. Eurocontrol, for instance, had said that they would not need such an aid as they already had their own internal rater training system.

The following tests were mentioned as beginning a list to be completed: Eurocontrol's ELPAC, RMIT's RELTA, the UKCAAi test, Mayflower's TEA test, ENAC's test, Anglo-Continental, Sofia University, GTELP, the New Zealand test, Versant, TELLCAP, Aerosolutions, a Brazilian test, the St. Petersburg and other Russian Federation tests. ASTAC had speech samples, but no test.

The question of whether only tests approved by the CAAs should be used was raised.

CA pointed out that the way a given speech sample had been rated in a particular test was not relevant for the present mission.

NB reminded the panel that ICAO had to inform the 190 contracting States of the sample gathering process. There would be a better response if ICAO announced that it was partnering with ICAEA. The disclaimers of what the RSS project would not do were essential, i.e. non-disclosure of sources and identities, non-jeopardy, non-accreditation etc. There would have to be a general communiqué (State letter) to which the test providers would be copied and a more detailed communication with test providers.

It was asked how many source languages should be used. In any case, both weak and strong candidates should be included for each language. Test providers should indicate language backgrounds, professional profiles and rough levels. Levels 3, 4 and 5 should be covered; Level 6 was too problematic and most tests did not provide enough evidence to assess Level 6.

All the other five official ICAO languages had to be represented (French, Spanish, Russian, Chinese and Arabic), plus Portuguese, Korean, Hungarian, Japanese, African languages, German or Scandinavian languages.

A list of things which would not be done and which would be done with the samples was drawn up:

- Not used for approval / endorsement / certification
- No bearing on licensing
- No individual feedback
- Not returned

The project would:

- Acknowledge participation
- Make candidate and test anonymous
- Possibly select extracts or tasks
- Provide the industry with a robust, wide-ranging aid
- Include as many characteristics (level, source language etc.) as possible
- Provide feedback freely through the website.

#### **1.4 Test materials and samples**

HE suggested that a form be drawn up for test providers to give information about a task / speech sample. He asked whether visuals should be included. There followed a discussion about the practicality and necessity of including the visual cues, such as photos, used in certain tasks.

SD was of the opinion that the focus should be on the language produced rather than the visual cues and other means. Af believed that the panel should concentrate on voice-only communication and that it would be more practical to eliminate visuals. BS pointed out that in RELTA visuals were only used for speech generation and not for the raters' use. It was generally agreed that asking providers for visuals, task descriptions and purpose was not practical. TG reminded the panel that the less work demanded of the providers the better it would be.

#### **1.5 Criteria and levels**

The question of which levels should be requested was raised. CA suggested that the panel express a 'particular interest' in Levels 3, 4 and 5 as by casting the net wide more samples should be collected.

Af said that a rating matrix incorporating raters, tasks, L1 language and levels needed to be developed.

NB said that getting more raters involved would help raise the bar across the industry. As a by-product, better rater training would be encouraged. In the process, raters would become aware of inconsistencies. Af agreed that it would be an opportunity to re-establish the standard.

GS thought that the key focus was on the difference between Levels 3 and 4.

It was decided to address the criteria of sample selection later in the process once the test providers had been contacted, so Items 1.6 to 1.12 were to be addressed at a later date once the providers had been contacted and the first samples had been received and screened.

## **2. Raters**

Af believed that raters should be given advance warning. TG wondered who the raters were and how they could be contacted. TG thought that test providers would object to receiving rater guidance and that communication with raters had to be diplomatic.

It was generally agreed that the rating should be of particular tasks and not of a test.

NB reminded the panel that the RSS process was not a judgement of a particular test. She raised the question of what was in it for the raters who took part. It would be an opportunity to conduct an internal calibration. TG pointed out that raters should not rate tests they work on themselves; this had to be included in the matrix.

CA proposed that the first step was to contact providers about the project in general, first of all through the ICAO State letter. Then the providers and CAAs should be contacted to identify raters while the initial samples were being screened by the panel.

### **2.1, 2.2 Criteria for rater selection and determining language proficiency**

Circular 318 (6.4.1) stated that raters should have a minimum of ICAO Level 5. It was pointed out, however, that raters were not using the language in an operational context and that the criteria of the ICAO Rating Scale were not relevant or only indicative.

There was an extended discussion about the requisite level of raters and how this could be verified, whether raters should assess themselves, whether the panel should request TOEFL, IELTS, TOEIC scores etc. HE reminded the panel that there were many raters operating freelance outside particular tests.

NB suggested that the raters should simply be asked whether they were Level 5 and provided with a blank field for further information or relevant experience.

Af believed that data analysis would eliminate the outliers and that it was better to cast the net wide, so involving more raters in the process and engendering a greater formative effect. The screening process for raters to eliminate unreliable raters would provide much useful data on current practice to facilitate improvement.

PS informed the panel that ICAEA intended to provide a feedback mechanism within the RSS website for those participating in the project to be able to give feedback and experience on issues such as the interpretation and wording of the rating scale, LPR implementation etc. This input would be collated and edited by Donna L'Hôte of the Centre de Linguistique Appliquée of Besançon University and published at the end of the

RSS project in order to contribute to longer term enhancements to the overall LPR implementation framework.

It was pointed out that many raters worked directly for CAAs and FTOs.

SD remarked that he was not sure that Thai raters would cooperate as there were only about ten raters operating in Thailand and they all worked for the Thai CAA.

Raters would be able to register on line either at the dedicated RSS website or via the specific pages of the ICAO or ICAEA websites.

It was suggested that CANSO, IATA etc. be contacted for assistance in distributing information about the RSS project.

## **2.4 Data collection and IT infrastructure**

The design of the data collection process would facilitate Ute's task.

BS suggested that [www.icaea.pansa.pl](http://www.icaea.pansa.pl) could be used for rater registration.

NB said that ICAO could possibly offer a secure site if the functionalities required were not too sophisticated. She wondered whether site management could be delegated to someone outside ICAO. CA observed that a 'shadow' server on the ICAO site could redirect communications automatically to an external server, but that there were obvious advantages in operating under the ICAO logo.

Af said that Mike French could set up an independent site which would not be costly. It would have two sections: 1) for storage of audio files which could be streamed, and 2) for instructions, data collection etc. PS would meet with Mike to discuss design specifications. Mike would need an IT contact at ICAO.

Good security was a prerequisite.

There would be a function for feedback gathering and a follow-up questionnaire once the rating had been performed.

It was agreed that the design of the website had to offer selective display of specific skills, overall levels, task by task, 1<sup>st</sup> language of the speaker etc.

After some discussion, it was agreed that transcriptions were not necessary and that the rationales should follow a single matrix (e.g. reference to descriptors and three quoted examples), have a consistent style and be simpler than those drafted in the first edition of the training aid.

HE updated RVB by telephone on the morning's proceedings and RVB said he would be happy to listen to samples and would take part in the next RSS panel meeting planned for March 2010.

There was some discussion over the amount of work which could be expected of the raters. The working hypothesis retained was 10 hours over 2 months.

Although the number of samples which would be received was considered to be a totally unknown quantity, the processing of sixty samples by the panel was adopted as a plausible figure, although two hundred was not impossible. BS and GS aired the view that the panel might not manage to obtain enough real samples and that a contingency plan by which samples were generated with different 'mock' test-takers using approved scenarios might have to be considered.

The panel discussed the overall time frame. The ICAO Council meeting in March 2010 (for announcing a realistic delivery date) and the final implementation date of 5 March 2011 were seen as the major milestones. It was agreed that the RSS product should be available for downloading as soon as possible in order to leave time for it to be of use before March 2011. Given the work involved, all agreed that anything much less than twelve months (October / November 2010) was not a realistic option and that perhaps Christmas 2010 should be proposed, leaving the possibility of completing ahead of time.

PS suggested that as the resulting training aid was to be delivered (free of charge) from the ICAO website, nothing prevented the group from delivering it incrementally, i.e. a small representative range of complete robust audio samples and rationales or a larger number of samples with only their scores. The remaining samples / rationales could be uploaded gradually. In theory, this approach would allow some samples to be available from the summer of 2010. This would have the added value of stimulating and maintaining interest in the website. In either case, the introductory and reference texts would need to accompany the samples.

The panel developed a draft action plan as follows:

### RSS Action Plan

Date	Action	Responsibility	Status / Notes
31.10.09	Prepare a draft ICAO State letter to inform CAAs of the joint ICAO / ICAEA RSS project	NB, PS	Done
02.11.09	Draft minutes of 30.10.09 meeting	PS	Done
04.11.09	Meet with Mike French and discuss the design specifications of the RSS website and report back to the panel	PS	Done
20.11.09	ICAO approval of draft State letter and signature by Secretary General	NB	
23.11.09	Second follow-up letter to CAAs, test providers, airlines soliciting participation in the project by providing speech samples	HE, TG	
01.12.09	Information page and FAQs to appear on <a href="http://www.icaea.pansa.pl">www.icaea.pansa.pl</a>	HE, BS, Af, PS	

31.12.09	Data base website operational to receive samples	Mike ffrench, HE, PS	
31.12.09	Screening checklist	Af, HE, PS	
15.01.10	Deadline for test providers to indicate their willingness to participate		
15.01.10	Reminder to test providers and request for the contact of accredited raters	HE, TG	
31.01.10	Letter to raters and rater guidelines	HE, TG	
31.01.10	Fields and information for raters available on RSS website	Mike ffrench, HE, PS	
15.02.10	First deadline to receive samples		
15-28.02.10	Pre-screening of audio samples received using checklist	led by HE	
28.02.10	Deadline for raters' registration		
01-20.03.10	Final screening of and comments on samples retained	4 linguists 4 SMEs	
x.03.10	Announce scheduled delivery dates to ICAO Council	NB	
22-26.03.10	3-day design meeting of RSS panel (UK or Warsaw) looking at raters, selecting samples, assigning samples to raters, developing Matrix (Af)		
12.04-31.05.10	First wave of rating		
30.04.10	Draft introductory textual framework for training aid	Af, PS	
31.05.10	Reminder to raters who have not transmitted ratings	HE, TG	
31.05-30.06.10	Second wave of rating		
15.06-15.07.10	First data analysis	UK	
15.07-15.08.10	Second data analysis	Lancaster	
31.08.10	First upload and delivery of rated speech samples from ICAO website (samples, textual framework and scores only)	NB, Mike ffrench, HE, PS	
01.09-30.11.10	Editing of speech sample rationales	Af and team	
15.09.10	Invitation to raters to complete questionnaire and feedback	HE, TG	
30.09.10	Second upload and delivery of rated speech samples from ICAO website (with first batch of rationales)	NB, Mike ffrench, HE, PS	
11.10-15.11.10	Collation and editing of raters' feedback and questionnaires	Donna L'Hôte	
31.10.10	Third upload and delivery of rated speech samples from ICAO website	NB, Mike ffrench, HE, PS	
17.12.10	Final upload and delivery of rated speech samples from ICAO website	NB, Mike ffrench, HE, PS	
17.12.10	Report on rater feedback delivered	Donna L'Hôte	
24.12.10	RSS projected completed		

Some minor modifications were made to the specifications document and the projected schedule updated.

PS reminded the panel that ICAEA was fully committed to the RSS project and was funding its launch, but had not enough internal resources to ensure its completion. Some external sponsorship and funding would be necessary. Obtaining such sponsorship would be easier once ICAO had officially announced its partnership with ICAEA on this project.

CA indicated that Lancaster University would be able to supply some support with administrative tasks. This offer was greatly appreciated.

The panel agreed to communicate by group e-mail. There was considerable satisfaction about the amount of ground covered and the practical outcomes of the meeting which exceeded the panel's expectations. Moreover, it had become clear that conducting the RSS project would have other positive side-effects which would enhance standards of rating practice and provide the community with important information. The panel agreed to meet again in Europe during the last week in March 2010.

The meeting ended at 3.15 p.m.