

## PLENARY SESSION 2

**Magdalena Večeřová**, Language testing consultant and expert to Eurocontrol

### BIO DATA

*After receiving her master's degree (MA) in English literature and philology Magdalena started her career as a language teacher in the Czech Armed Forces. She also studied at the Canadian Language Training Faculty in Ottawa where she specialised in ESP (English for Specific Purposes). Soon after her return from Canada she became a member of a new language testing team for the Czech military and was sent to the University of Lancaster in Great Britain where she completed another MA in language testing. She also co-operated with the Peacekeeping English Project of the British Council and provided training and workshops in language testing for participants from more than 20 countries (both in the Czech Republic and abroad).*

*In 2005 she started her co-operation with EUROCONTROL Luxembourg where she was involved in the development of the ELPAC test. Magdalena currently works as an independent language testing consultant and is involved in both the ELPAC test and a new test for student controllers which is being developed by EUROCONTROL.*

### **“Test evaluation: Do our tests fulfil the ICAO requirements?” with Q & A**

#### SUMMARY

*Testing language proficiency in aviation is of very high stakes. Unreliable and invalid tests compromise safety and thus service providers, operators and regulators should make every effort to ensure that a valid and reliable test is being used and standards are being kept. Best practice has to be applied not only in the test development process but also in test implementation.*

*There are currently many language tests claiming that they meet the ICAO requirements and assess language proficiency according to the ICAO scale. This is, unfortunately, not always the case. This paper and the accompanying presentation provide practical advice on how to evaluate language tests and what to do when the test being used does not meet the ICAO requirements.*

#### **1. Introduction**

As of March 5th 2008 air traffic controllers and pilots operating in internationally designated airspace and on international air routes have to demonstrate their proficiency in the language(s) they use for aeronautical communication. Level 4 of ICAO's language proficiency requirements is set as the operational standard which air traffic controllers and pilots must meet.

Air Navigation Service Providers, Aircraft Operators and National Regulatory Authorities should have tested all their personnel and taken appropriate measures where level 4 has not been achieved, or have in place an implementation plan to be published and completed by March 5<sup>th</sup> 2011 (ICAO Resolution A36-11, October 2007).

ICAO Doc 9835 (Manual on the Implementation of ICAO Language Proficiency Requirements), chapter 6 states: “*Reliable, effective, legitimate testing systems are required to ensure that pilots and controllers have adequate levels of English language proficiency. ...language testing for licensing purposes needs to be of highest calibre...*”.

The information published on the ICAO website (<http://www.icao.int/fsix>) shows a very different picture from what was originally expected. Many States have not even started testing and many are using unreliable and invalid tests. All examples used in this paper are based on information from State Letters published on the ICAO website.

## **2. Test evaluation**

There are currently many codes of practice providing guidance on how to develop, implement and evaluate language tests. It is beyond the scope of this paper to discuss the test evaluation in detail and thus only the main principles will be explained. More information can be obtained from the links provided at the end of this paper.

### **2.1. Test purpose**

The ICAO Language Proficiency Requirements (LPR) require the use of specific purpose language proficiency tests. The test content must be based on a needs analysis and test tasks have to be representative of the target language situations.

The data published on the ICAO website are incomplete as many States have not yet submitted the required information. Yet there are several instances where tests being used do not meet the requirements.

#### Example 1:

*OPI (Oral Proficiency Interview) is used to assess the language proficiency of pilots / controllers.*

Problem: A perfect example of a test misuse. OPI was not designed to test the language proficiency of pilots and controllers and was validated for a different purpose. There is no aviation element. Results from this test bear no reference to the ICAO scale and are thus meaningless as they cannot be translated into ICAO levels.

Solution: A new test has to be used. It must be trialled on a representative sample of the target population and test tasks have to be representative of the target language situations.

*(Editor’s note: This topic has been expanded upon in several exchanges by the author and other contributors on the icaea\_world forum in May 2008)*

#### Example 2:

*Diagnostic instead of proficiency test being used.*

Problem: No further information provided and thus it is difficult to comment on the test. In general, diagnostic tests are used to identify areas where a test taker needs help.

Solution: Results from this test could be used by teachers to prepare remedial language programmes. A specific purpose language proficiency test has to be used to evaluate the language proficiency.

#### Example 3:

*Placement tests and informal interviews being used.*

Problem: Placement tests assess a student’s level in order to place him or her in an appropriate language course. Informal interviews might give an indication of someone’s level, nothing more. Test versions should be parallel and this cannot be the case with informal interviews. Thus each candidate receives a different test.

Solution: Similar as above. Use the information obtained from the two tests to place test takers into various language classes. For licensing purposes use a specific purpose language proficiency test that meets the ICAO LPR.

## **2.2. Reliability**

Reliability refers to the degree to which test scores are free from different types of chance effects. No test will achieve a perfect reliability (1.0) but in high stakes testing tests with reliability coefficients as close to 1.0 as possible should be used. In speaking tests intra- and inter-rater reliability should be continually checked. When inconsistencies appear the raters should undergo refresher training and be re-accredited upon successful completion of the training. Examples of problematic tests are presented below:

### Example 1:

*The test developer stated that tests were trialled and the results showed that the test was reliable.*

Problem: If no further information is provided then the quality of the test is in doubt. Each statement has to be supported by evidence. Also, it is not enough to say that the trialled test version was reliable. The reliability of the test must be monitored during all test administrations.

Solution: There may be nothing wrong with the test itself. The problem might be on the side of the test provider who fails to communicate the necessary information. Request evidence to support the statements about the test reliability and information on how the reliability was and is computed.

## **2.3. Validity**

Validity is a rather complex test characteristic and the scope of this paper does not enable detailed discussion. In general, a test is valid when it measures what it is supposed to measure. Validity pertains to the correctness of the inferences or decisions made on the basis of test scores. Therefore each test version should be trialled on a representative sample of the target population. Test items/tasks must represent the language skills needed in the specific content domain. Examples of several problematic tests currently being used are presented below:

### Example 1:

*Controllers are assessed in listening, speaking and reading.*

Problem: ICAO rating scale includes listening comprehension and speaking. Testing reading does not seem to represent the language skills needed by controllers in their work. Levels assigned cannot be ICAO levels as there is no scale for reading comprehension. A perfect example of construct irrelevance.

Solution: If the listening and speaking tests follow all ICAO requirements on language proficiency testing they should be retained and used. The reading comprehension test should not be used for licensing purposes.

### Example 2:

*A speaking test (face to face) assessing all 6 ICAO levels. The test lasts approximately 20 minutes and there is one interlocutor and one rater.*

Problems: Listening comprehension should be "...assessed separately from speaking through an objective assessment tool. This is necessary to reduce the impact of variables associated

*with oral proficiency interfering with assessment of proficiency in listening comprehension”* (Michael Kay: Test Evaluation Criteria).

Also it would be impossible to assess if the test taker has the ability "...to comprehend a range of speech varieties (dialect and/or accents) or registers." (Comprehension: Level 5 on the ICAO rating scale). Level 6 goes far beyond the scope of aeronautical communications (vocabulary is idiomatic, nuanced and sensitive to register, level 6 speaker varies speech flow for stylistic effect, demonstrates comprehension of linguistic and cultural subtleties, is sensitive to verbal and non-verbal cues and responds to them appropriately, etc.). A 20-minute test cannot include tasks which would test language at all 6 ICAO levels in a valid and reliable way. This test seems to be a perfect example of construct under-representation.

Solution: Evaluate the speaking test thoroughly to see to what extent it is capable of eliciting a rateable sample of language. Review all test items / tasks and prepare new ones. Test tasks have to be representative of the target language situations. A separate listening comprehension test has to be developed.

#### Example 3:

*Informal listening and speaking tests being used.*

Problem: There is no more information provided but the word “informal” seems to suggest that the tests have not been validated and might be just general English language tests. They are probably unstructured and versions are not parallel. If this is the case then the results cannot be translated into ICAO levels.

Solution: Evaluate the test using the guidelines provided by ICAO and if the test does not conform to the guidelines do not use it for licensing purposes.

#### Example 4:

*Oral interaction consisting of introduction, communication on common topics, auditing an aviation theme, auditing a real radio communication, testing task simulation.*

Problem: The description of the test is not very clear. No sample test was provided and thus it is very difficult to see what the test takers actually have to do. Again, no listening comprehension test is included and thus it is another example of construct under-representation.

Solution: Evaluate the test using the guidelines provided by ICAO and if the test does not conform to the guidelines do not use it for licensing purposes.

#### Example 5:

*Language proficiency test in plain English.*

Problem: Radiotelephony is not plain English only. Phraseology is an integral part of pilot/controller communication. Such a test cannot assess language proficiency according to the ICAO scale.

Solution: This test cannot be used for licensing purposes.

### **2.4. Fairness**

Each test should be fair to all candidates, regardless of their gender, age, language background, race, ATC rating or the aircraft type rating a pilot has. Each test version must be trialled on a representative sample of the target population and bias analyses have to be conducted to make sure that the test does not disadvantage a particular group of candidates. An example of a problematic test is presented below:

#### Example:

*Test being used is a “sample interview and estimated grading”.*

**Problem:** It seems that the test being used has not been trialled and does not use the ICAO scale. A test should measure, not estimate. Test versions are probably not parallel and thus it is impossible to see whether and to what extent could some candidates be disadvantaged.

**Solution:** Use the results from the test for placement purposes and develop (or buy) a new test to use for licensing purposes.

### **3. General considerations**

There are many States who until now (May 2008) have not provided any information on their compliance. Some, on the other hand, use “grandfather rights” to declare that all pilots or controllers are at level 4. Some state that as they have had no major incidents where lack of language skills would be a contributory factor, everybody must be at level 4 or above. This is an extremely dangerous approach. Nothing should be taken for granted and awarding levels of language proficiency must be based on measurement, not on assumptions.

### **4. Basic guidelines for evaluation of language tests to meet the ICAO LPR**

- There should be separate tests of listening comprehension and speaking.
- Specific purpose language proficiency tests that meet ICAO LPR should be used (not diagnostic, placement, progress or plain language tests).
- Needs analysis has to be conducted.
- All test versions have to be trialled on a representative sample of the target population.
- Test items / tasks have to be representative of language skills needed in the content domain (radiotelephony). Different tests should be used for pilots and controllers.
- ICAO rating scale has to be used for assessment.
- Tests have to be sufficiently reliable to permit stable measurement of test takers' abilities.
- Tests are being validated and periodically reviewed.
- Tests are being sustained: new test versions are developed and trialled, testing standards (administration, interlocuting, assessing, ...) are monitored and maintained.
- There exist comprehensive and clear guidelines for everybody involved in test administration.

Detailed guidelines for test design and evaluation can be found in the following links:

#### **General Guidelines:**

- ALTE Principles of Good Practice: [http://www.alte.org/quality\\_assurance/index.php](http://www.alte.org/quality_assurance/index.php)
- Code of Ethics for ILTA: <http://www.iltaonline.com/code.pdf>
- ILTA Code of Practice: <http://www.iltaonline.com/ILTA-COP-ver3-21Jun2006.pdf>
- The Standards for Educational and Psychological Testing:  
<http://www.apa.org/science/standards.html>
- Code of Fair Testing Practices in Education: <http://www.apa.org/science/FinalCode.pdf>
- EALTA Guidelines for Good Practice in Language Testing and Assessment  
<http://www.ealta.eu.org/documents/archive/guidelines/English.pdf>

#### **Specific Guidelines**

- <http://www.icao.int/fsix/lp.cfm>
- ICAO Doc 9835 (Manual on the Implementation of ICAO Language Proficiency Requirements),

- EANPG 48\_Report\_Appendix J: Recommended qualifications for raters of tests to meet ICAO LPR
- EANPG 48\_Report\_Appendix K: Recommended qualifications for interlocutors of tests to meet ICAO LPR
- Test Evaluation Criteria prepared by Michael Kay (ICAEA Board member)
- Selecting a language proficiency test to meet ICAO LPR (presented by Magdalena Vecerova at IALS/2)

### ***Conclusion***

Language testing in aviation is of very high stakes but not all service providers, operators and regulators seem to be aware of this fact. It is important to make every effort to ensure that valid and reliable tests are used. Unreliable and invalid tests compromise safety in aviation and this is what we all should want to avoid.