

Selecting a language proficiency test to meet ICAO language proficiency requirements

Magdalena Vecerova
EUROCONTROL (European Organisation for the Safety of Air Navigation)
Luxembourg

Summary

The paper will guide the service providers, operators and regulators through the test selection process and explain what should be taken into consideration and why. It will present not only a series of questions to be asked but also the evidence that one should seek.

Introduction

This paper is aimed at Air Navigation Service Providers, Aircraft Operators and National Regulatory Authorities who have yet to select a language proficiency test to meet the ICAO language proficiency requirements and who feel they do not have the required expertise to do so. Due to the limitations on the length of the paper, it is not possible to provide a comprehensive guide to the test selection process. The paper does, nevertheless, provide a list of the most important issues which have to be taken into consideration when selecting a language test.

Selecting a test

A perfect test does not exist. Each test has its limitations and there is no test which would suit all. Therefore, the decision makers must first study the ICAO language proficiency requirements and then examine their own situation (available resources) and based on their findings choose a test which meets their requirements in the best possible way. The most important information to bear in mind is that each test should be used only for the purpose it was designed for, otherwise the inferences one makes based on the test results are not valid. Choosing a test, which would not measure language proficiency in a valid and reliable way, would compromise safety in aviation and thus decision makers should select a test with the greatest care possible.

Testing language proficiency according to the ICAO rating scales means that specific purpose language proficiency tests have to be developed and used. Thus the test content must be based on a needs analysis and test tasks have to be representative of the target language situations. There are a number of universal questions one should ask test developers. All test developers must provide comprehensive answers and if they decide not to, it is the first sign of possible problems with their test. Before you decide to use a test examine its quality and sustainability carefully. Tests with incomplete documentation, or tests which you cannot evaluate thoroughly, should be excluded immediately.

Areas to cover in the test evaluation process

Each test developer should provide information about the following: *Test purpose, description of the test development process, technical information (reliability, validity, fairness ...), sustainability, required resources for test administration.*

Test purpose

The basic question to be asked is who the test is intended for and what it is supposed to measure. Thus the test developer must describe the test population (pilots, air traffic controllers, ...) and explain how the test should be used. Also, inappropriate test usages should be identified. Two examples of inappropriate test usages are presented below.

Example 1: *There is a test measuring language proficiency at ICAO level 4 and it was designed for pilots. The test developer says that the test reflects the job related context of pilots and it tests only language proficiency, not knowledge of procedures or any other technical knowledge. There are two test papers – listening comprehension and oral interaction. The test developers say that the test has been validated for the intended purpose – testing language proficiency of pilots at ICAO level 4. A regulator sees this test and likes it. But he wants a test for controllers too so decides to use the same test for both pilots and controllers. This regulator also wants to test ICAO level 5. So the final result is this: only the oral interaction paper will be used as when talking one must also listen and thus listening will also be assessed. The examiners will assess levels 4 and 5, level 5 being simply somebody really good. And both pilots and controllers will be tested with the same test so nobody should complain. Or should they?*

Mistake #1 - Using the test for both pilots and controllers: Needs analysis for a test for controllers had not been conducted, the language functions that controllers use had not been identified and the job related context might be actually quite different from the one identified for pilots. The test had not been trialled on controllers and there is absolutely no evidence that the test is suitable for this purpose. Thus, controllers might be failing the test not because of their poor language abilities but because of other factors.

Mistake #2 - Using only one part of the test: The test was designed in two parts and was trialled and validated in this format. Using just one part of the test means a serious misuse of the test and thus interpretations of test results and decisions based on them are invalid.

Mistake #3 - Assessing different language proficiency levels: The test has no tasks at ICAO level 5 as all tasks were trialled to test level 4. The fact that someone might actually be at ICAO level 5 is not measurable by the test. Only a test which has items and tasks at level 5 and is capable of eliciting language at level 5 can be used to measure this level. When a candidate does very well at a level 4 test there is no guarantee that he would pass a level 5 test.

Example 2: *A country has been using a placement test for a few years and it has yielded satisfactory results. At least that is what the teachers said. When they wanted to sort out people into groups with similar language abilities, the test results were fairly accurate. And in cases where students proved to be above or below the level of their colleagues they were simply transferred to a different group. The test has two parts – multiple choice grammar test and a general English listening comprehension test. So, why bother to look for another test when we have one already?*

Mistake #1 - Using a placement test instead of a proficiency test: The purpose of the placement test is to sort out students into groups for teaching purposes. Placement tests are usually not of high stakes, unlike proficiency tests where, based on test results, very important decisions are being made. Every test has to be used only for the purpose it was designed and validated for, all other usages are invalid.

Mistake #2 - The test is not testing listening and speaking in the job related context: There is a listening comprehension test but only in general English, which may be checking whether a candidate understands, people who talk about their holiday plans – nothing to do with aviation English. Speaking ability is not assessed at all.

Mistake #3 - The test does not assess proficiency according to the ICAO scale: The test may sort out students into different groups but there is no information on how it is done and what the different levels really mean. ICAO levels refer to job related context so this test must be using a different scale.

Test development

Decision makers should require information on who was involved in the test development. In the case of tests developed to meet ICAO language proficiency requirements, both English language experts (linguistics background) and subject matter experts (controllers, pilots, ...) should be involved. The development team has to have language testing experts and also statistician(s). Ask for a list of the qualifications and expertise of all people involved in the development. The fact that somebody speaks excellent English does not automatically make this person a good language test developer. Also, there has to be a team working on the test, not one or two people.

Decision makers should demand information on how the test items and tasks were developed, trialled and analysed. Each test item or a task has to be trialled on a representative sample of the target population, otherwise inferences made on the basis of test scores would not be valid. All test versions must be of the same level of difficulty and have to sample the content domain adequately. Also all candidates should receive the “same test” no matter which test version they take or who is examining them. If there is a test where one version is easier than another or when some examiners are more lenient than others then such a test is neither reliable nor valid and should not be used at all.

Technical information (reliability, validity, fairness)

Test developers might impress you with complex statistical reports, which, if you are not a statistician, you might not be able to understand. If you cannot consult an expert, this is the basic information you should be looking for:

Reliability

Reliability is often defined as a consistency of measurement and refers to the degree to which test scores are free from different types of chance effects. It is usually reported in the form of a reliability coefficient which can range from 0.0 to 1.0. The higher the figure, the better. No test will achieve a perfect reliability (1.0) but one should look for tests with reliability coefficients as close to 1.0 as possible. When evaluating speaking tests one should also make sure that examiners (assessors / raters) are as consistent as possible. There are two kinds of reliability one should be aware of:

- *Intra-rater reliability:* The extent to which a particular rater is consistent in using a proficiency scale.
- *Inter-rater reliability:* The level of consensus between two or more independent raters in their judgement of candidates' performance.

Validity

Validity is a rather complex test characteristic but when simplified we could say that the test is valid when it measures what it is supposed to measure. And, above all, validity pertains to the correctness of the inferences or decisions made on the basis of test scores. As has already been said, each test can be valid only for the purpose it was designed for. There are many important questions one should ask.

Request evidence that the test has been trialled on a representative sample of the target population and that the sample was big enough to generalise from. Ask how the difficulty levels of items were established. It is easy to say that the test measures language proficiency at ICAO level 4 but how does one know? When choosing the trialling population, one has to make sure that there are candidates with different ICAO proficiency levels and the information on their proficiency levels needs to be available and be measured with a validated instrument.

Test developers should also provide information on how the cut-off score (score required to pass) was set. Is it simply 75% because that is the tradition? Or has a more scientific approach been taken and statistical analyses and expert standard setting conducted? One should of course prefer the latter, but still require a detailed description of the process of arriving at the cut-off score. It is easy to say something was done, so never forget to require evidence that it is really the case.

Language proficiency testing in aviation is of very high stakes and the tests should reflect the job-related context of the candidates. Needs analysis has to be conducted and a detailed description of the test content provided. Test items/tasks must represent the language skills needed in the specific content domain. One of the basic requirements is involvement of subject matter experts in the test development and an independent expert evaluation of the test.

Tests are, of course, only samples from a content domain. One has to make sure that the test is as representative of the content domain as possible. If a test omits one or more important aspects of the content domain, we talk about “construct under-representation”. A simple example might be a test for controllers where they would have a test of phraseology and then a short interview about what they like about their job. If a test contains items or tasks which test something else than what the test is supposed to measure we talk about “construct irrelevance”. A simple example might be a test of listening comprehension with multiple-choice questions where the options are so long and in such a complex language that candidates fail to understand the questions. Then it is not their listening ability which is being measured, but their ability to understand long and complex multiple-choice questions.

Fairness

Each test should be fair to all candidates, regardless of their gender, language background, race, ATC rating or the aircraft type rating a pilot has. Request evidence that the test was trialled on a representative sample of the target population that bias analyses have been conducted and the test does not disadvantage a particular group of candidates. Two examples of problematic tests are presented below:

Example 1: *There is a test of listening comprehension for controllers and the reported reliability coefficient is 0.98. The test was developed by one language teacher who has been teaching aviation English for over 20 years, The test was trialled on 50 (20 TWR, 30 APP) controllers working in one ATC unit.*

Mistake #1 - No development team: The test was developed by one person only. No matter how good somebody might be, there always has to be a team of experts involved. Also, no subject matter expert (i.e. air traffic controller) was involved.

Mistake #2 - Trialling population not large enough and not representative of the target population: The test was not trialled on a representative sample of the target language population. If it is intended to measure language proficiency of controllers, controllers with all ATC functions should be involved in the trial so that bias analysis (ATC function, language background, etc.) can be conducted.

Mistake #3 - Believing that a high reliability coefficient solves all problems: It is true that each test should have the reliability coefficient as close to 1.0 as possible. It should be as a consistent measure as possible but it also has to measure what it is supposed to measure. A test can be perfectly reliable yet not valid.

Example 2: *Pilots in one airline company are to be tested during their line check. The rationale being that they speak English anyway so one can test two things at the same time.*

Mistake #1 - Test misuse / different test purpose: The line check has never been designed to measure language proficiency and has never been validated for this purpose.

Mistake #2 - Construct under-representation: The content domain would definitely not be sampled adequately as important parts of the content domain would be ignored.

Mistake #3 - Test versions not being parallel: The test would be different for different pilots as it is unpredictable what will happen during a flight. One pilot might use standard phraseology almost all the time, never experiencing anything unusual and another might go through an emergency or unusual situation.

Sustainability

Test developers must provide information on how the test will be sustained. Require information on how new test versions will be developed and trialled, how the test developers want to ensure that all test versions are parallel and how will the testing standards (administration, interlocuting, assessing, ...) be monitored and maintained.

Test administration

Resources (human, material, time) for test administration have to be specified. Test developers should also provide evidence that the test was trialled under and validated for the same conditions as intended for live test administration. Comprehensive and clear guidelines for everybody involved in test administration (including candidates) have to be provided.

Conclusion

Language testing in aviation is of very high stakes. Before making a decision on which test to use, service providers, operators and regulators should evaluate each test package thoroughly. It is important to make every effort to ensure that a valid and reliable test is chosen. If possible, language-testing experts should be used for test evaluation. In situations where there are no language testing experts available, this paper might serve as a guidance to what to look for when selecting a test and one should make sure that at least the basic issues covered in this paper are investigated. Unreliable and invalid tests compromise safety in aviation and this is what we all want to avoid.

References

Bachman L.F. and Palmer A.S. *Language Testing in Practice*. Oxford University Press (1996).

Douglas, D. *Assessing Languages for Specific Purposes*. Cambridge Language Assessment Series. Cambridge University Press (2000).

Rudner, L.M. (1994). *Questions to Ask When Evaluating Tests*. <http://pareonline.net/getvn.asp?v=4&n=2> (Practical Assessment, Research & Evaluation) (downloaded 28/04/06).

Le Vi-Nhuan and Klein S. P. *Technical Criteria for Evaluating Tests* http://192.5.14.110/pubs/monograph_reports/MR1554/MR1554.ch3.pdf (downloaded 16/03/07).

— — — — —