Why is language testing for the ICAO LPRs in such a poor state?

TWELVE SHORT ESSAYS

Henry Emery, June 2022

Contents

Introduction

Language testing for the ICAO Language Proficiency Requirements (LPRs) is exceptionally high-stakes. Test scores have a direct effect on individuals, organisations and safety. While the industry deserves the highest standards that the field of language testing has to offer, aviation English testing is in a very poor state. This collection of twelve short essays, first published as a series of articles on LinkedIn in April and May 2022, explores the issues and summarises my personal incredulity with the status quo. I know I am not alone, though I am surprised about how quiet the aviation community is about it.

Caveats:

1. I don't own the ideas which follow, but I am giving them a voice.
2. The problem is complex and multi-dimensional. There is more to it than is written here.
3. It's not my place to call out test service providers by name.
4. Good practice does exist, but it is the rare exception, not the rule.

1. Inappropriate regulatory guidance

There is a lack of language assessment literacy among regulators leading to inappropriate guidance on language assessment. This, from the FAA[1]:

"Read the introduction sections of the documents to the applicant ... then request that the applicant read a portion of the text, ask the applicant to explain what they heard, and request that they write down in their words what they heard and read ... this will determine whether or not the applicant can communicate with ATC."

Until regulators better understand language testing and promulgate standards that help ensure tests are fit for purpose, pilots and controllers will continue to take poorly-constructed language tests that fail to address aeronautical radiotelephony communication.

2. Lack of operationally relevant tests

There is a lack of assessment literacy among test service providers and regulators resulting in an absence of operationally relevant test instruments. This, from ICAO[2]:

[1] See: https://www.faa.gov/documentLibrary/media/Advisory_Circular/AC_60-28B.pdf

[2] ICAO (2010) *Manual on the implementation of ICAO Language Proficiency Requirements* (2nd ed.) Doc 9835 AN/453

"The sole object of ICAO language proficiency requirements is aeronautical radiotelephony communications."

On scanning the websites of the plethora of aviation English test service providers, I can find just two whose instruments contain tasks that explicitly and directly address 1) listening comprehension and 2) speaking in the context of RT communications. The rest do it either partially or not at all.

Until regulators and test service providers better understand language testing and ensure the provision of tests which are operationally relevant, pilots and controllers will continue to take poorly-constructed language tests that fail to address aeronautical radiotelephony communication.

3. Generic test design

There is a lack of assessment literacy among test service providers and regulators which has led to a prevalence of generic aviation English tests designed for both pilots and air traffic controllers.

Pilots and controllers share the radio, but they do different jobs and pursue different objectives. Consequently, the way they use language differs. This, from ICAO[3]:

"While pilots and controllers are communication partners, they approach the task from different perspectives, and therefore their communication differs in purpose and standpoint".

"Because of the high stakes involved, pilots and air traffic controllers deserve to be tested in a context similar to that in which they work. Test content should, therefore, be relevant to their work roles".

The purpose of English language testing in accordance with the ICAO LPRs is to make inferences about the ability of pilots and controllers to perform on-the-job language tasks in English. To make valid inferences, we need to elicit domain-specific language performances. It is not possible to elicit domain-specific language performances of different populations using the same generic test tasks.

Until regulators and test service providers abandon the flawed notion that generic test tasks can adequately capture the specific language use domains of different personnel, pilots and controllers will continue to take poorly-constructed language tests that fail to address aeronautical radiotelephony communication.

[3] Ibid.

4. Underrepresentation of listening comprehension

There is a lack of assessment literacy among test service providers and regulators which has led to the inadequate treatment of listening comprehension in aviation English tests.

Pilots and controllers do A LOT of listening on the radio. Listening comprehension is at least 50% of the communicative load. Listening is a complex process that results in a partial, unique, fleeting and invisible picture in the listener's mind. Because the listening process is not directly observable, we need carefully designed instruments which present:

- Plenty of pre-recorded text which is representative of radio communications; and
- A variety of tasks designed to tap listening specifically.

Good tests of listening comprehension take time to administer. The higher the level at which we aim to assess, the more comprehensive and therefore the longer our listening tests need to be.

On scanning the websites of test service providers, two issues are evident:

1. Many assess the ability of pilots and controllers to understand radiotelephony communications:

- With very limited comprehensible input; and/or
- Without reference to recordings of radiotelephony communications at all.

2. Many assess listening through performance in speaking tasks. This results in poor measurement because:

- Speaking tasks cannot present the quantity or quality of comprehensible input required to enable robust measures of listening; and
- What people say is influenced by many things, only one of which is what they understand.

Human rating of listening through a test taker's performance in tasks designed primarily to elicit speaking leads to indirect, muddied and error-prone measurement.

Until test service providers better understand the listening construct and offer tests which adequately address listening comprehension in the specified domain, pilots and controllers will continue to take poorly-constructed language tests that fail to address aeronautical radiotelephony communication.

5. Poor construct definition

Test service providers fail to adequately conceive of the construct of professional language use which results in weak test instruments.

To test well, a test instrument must be built on a detailed description of the characteristics of the target language use domain. This description - the ‘construct’ - is fundamental because language use is the product of underlying skills which are not directly observable. To make inferences about these skills in any meaningful way, we need a construct definition that is both theoretically sound and as comprehensive and precise as possible. ICAO Document 9835 and the Rating Scale are an essential part of the test developer’s toolkit, but they are not in themselves the test construct. Construct definition is the job of the test service provider. Without a well-defined construct, a test is just a stab in the dark.

Let’s briefly explore one aspect of the construct: Professional language use.

If our agreed mission is to measure the professional language use of pilots and controllers, then tests must be firmly rooted in the operational domain. This means we need to clearly describe, with reference to established theory, how tests treat the relationship between operational knowledge and language use.

You cannot speak "aviation English" without operational knowledge. If you present any test prompt to a pilot or a controller that is even remotely connected with their work, what you get back is inevitably a blend of the two. Separating them is not only impossible (unless we have a chat about sport or cake), it is also undesirable: a basic principle of specific purpose language testing holds that tasks "should allow for an interaction between language knowledge and specific purpose content knowledge"[4]. Therefore, in aviation language testing, operational knowledge is an implicit part of the construct. This should be evident in the way test service providers communicate about their tests and how they realise the construct in test tasks.

This from a range of European aviation English test service providers:

- *“[TEST] is not a test of operational knowledge"*
- *"[TEST] focus ... is not on operational procedures"*

And this last one is the best:

- *"Don't be afraid to say something with incorrect content, as long as your ICAO English is perfect, there's nothing to worry!" [sic]*

[4] Douglas, D. (2000) *Assessing Languages for Specific Purposes* (CUP)

To be clear, language testers have no business directly assessing the accuracy of technical subject matter presented by test takers in language tests. But operational knowledge will always be present in a test taker's language performance in any test of "aviation English", even in tests which are weak representations of the construct of aeronautical communication.

The mission of aviation language testing is not to assess 'language proficiency'. It is to assess 'operational language proficiency', i.e. the language pilots and controllers need to do their jobs safely. The closer we get to authentic interaction between operational knowledge and language proficiency in test tasks, the better our tests will be. Undermining the central role of operational knowledge indicates a failure to adequately conceive of the professional language use construct with a corresponding failure to operationalise the construct in test tasks.

Consider the difference between:

- “[TEST] is not a test of operational knowledge"; and
- “[TEST] assumes that candidates have professional knowledge of aeronautical operations and procedures and standard radiotelephony phraseology. [TEST] engages but does not directly assess this knowledge.”

There is a subtle yet powerful distinction between these two descriptions. The first (real-world) description is misleading. The second (hypothetical) description positions operational knowledge as a central (if not explicitly assessed) part of the construct which opens the door to the design of test tasks that directly address operational language use.

Until test service providers better understand the professional language use construct and develop more theoretically-sound test instruments that reflect the operational language use domain, pilots and controllers will continue to take poorly-constructed language tests that fail to address aeronautical radiotelephony communication.

6. The regulatory and commercial backdrop

One reason language testing for the LPRs is in such a poor state is the regulatory and commercial backdrop.

Designing, maintaining, validating and administering a quality language test requires field-specific expertise and is costly and time-consuming. To recover costs, test service providers need either substantial volume or subsidies.

Compared to testing in other contexts, for example, in academic English where a small number of professional testing organisations assess millions of test takers each year, the total worldwide population of pilots and controllers is very small. Remove those with level 6 completely. Remove those with level 5 for 6 years. Remove those with level 4 for 3 years. What you are left with is a tiny annual candidature. Given the plethora of inadequately regulated commercial test service providers competing for this very small population, it is no wonder we see such poor quality.

To be clear: It is possible to provide quality aviation language testing. And it is possible to make money from aviation language testing. But in the current regulatory landscape, you can't do both. The numbers don't add up.

Non-commercial aviation language testing services require generous subsidies. Some are well funded and provide good services, but more often than not, non-commercial test service providers are run on a shoestring by passionate people who are obliged to operate without adequate training and support. To quote a phrase: "We, the willing, led by the unknowing, are doing the impossible for the ungrateful".

Until:

- Regulators implement approval procedures which create the conditions for quality commercial test service providers to thrive; and
- Non-commercial aviation language testing is supported with adequate training and funding;

Pilots and controllers will continue to take poorly-constructed language tests that fail to address aeronautical radiotelephony communication.

7. Lack of accountability

There is a complete lack of accountability in aviation language testing. This from EASA and ICAO:

- *"The assessment documentation should include at least … documentation demonstrating the assessment validity ... and reliability".*[5]
- *"Test service providers should supply documented evidence of the validity and reliability of their testing methods".*[6]

This, from experts in language testing:

[5] EASA (2016) Annex 1 - Part FCL (FCL.055)
[6] ICAO (2010) *Manual on the implementation of ICAO Language Proficiency Requirements* (2nd ed.) Doc 9835 AN/453

- *"The need to estimate and report information about reliability and measurement error … is explicitly stated in the professional standards for language testers. [It] is not only a matter of good testing practice; it is also a professional responsibility".*[7]
- *"The issue of accountability is important, and we would argue that certain minimum information ought to be made publicly available".*[8]

On scanning the websites of the dozens of test service providers in the marketplace, none offers any meaningful evidence for the validity and reliability of their instrument.

Until we have accountability in aviation language testing, pilots and controllers will continue to take poorly-constructed language tests that fail to address aeronautical radiotelephony communication.

8. Lack of industry-recognised services

There is a lack of industry-recognised testing services.

Evaluating a language test requires field-specific expertise which regulators typically don't have. Consequently, regulators approve sub-standard tests. By and large, national regulatory approval is no indicator of language test quality.

In 2010, ICAO acknowledged that 'many States still lack the expertise and resources to implement ICAO Guidance regarding the selection and development of appropriate testing tools'[9]. In 2012, with aviation and language industry partners, ICAO went on to launch the *Aviation English Language Testing Service*[10] with the goal of creating 'a pool of testing systems from which States can choose'. It was a brilliant concept, but it has sadly failed: of the dozens of aviation English tests that are available today, just one is recognised by ICAO. Why?

In the process of applying for ICAO recognition, the Test Service Provider (TSP) is required to submit evidence for the validity and reliability of their test instrument. The evidence requested[11] is perfectly reasonable. Any professional language assessment organisation will have some if not all of this evidence to hand. The problem is that the vast majority of TSPs that apply for ICAO recognition fail because they cannot provide this evidence. This is because:

[7] Alderson, J.C., Clapham, C. & Wall, D. (1995) *Language Test Construction and Evaluation* (CUP)
[8] Bachman, L. (2004) *Statistical Analyses for Language Testers* (CUP)
[9] ICAO (2010) *Language Proficiency: New Test Endorsement Process* ICAO Journal, 65:4, 30-31
[10] See: https://www4.icao.int/aelts/Home/RecognizedTests
[11] See: https://www4.icao.int/aelts/uploads/a%20guide%20to%20submitting%20validity%20evidence.pdf

- Gathering this evidence requires test trialling and analysis which is prohibitively costly (the $5,000 fee for ICAO recognition is trivial in comparison); and/or
- Gathering this evidence requires field-specific expertise which the TSP does not have; and/or
- The test instrument is fundamentally faulty which makes the gathering of evidence a pointless exercise.

This is why so few tests have achieved ICAO recognition, and explains, in part, the complete absence of accountability in aviation language testing.

Although ICAO recognition was established specifically to help authorities select and approve language tests on the basis of quality, it has unfortunately never gained traction. Instead, we see weak approval procedures at the national level. When national regulatory approval is cheap and easy to achieve, why would any TSP volunteer to jump through hoops with ICAO? In Europe, the situation is compounded: a weak approval of a weak test in one EASA member state grants the TSP access to markets across all EASA member states.

Some have been critical of ICAO recognition. Of course, the process is not perfect, but it could and should be a powerful tool for regulating aviation language testing. The reasons ICAO recognition has never had the desired impact are complex and multi-dimensional, but beyond political pressures and commercial challenges, they can be boiled down to two things: 1) poor assessment literacy amongst regulators leading to 2) poor assessment practice. Consequently, pilots and controllers continue to take poorly-constructed language tests that fail to address aeronautical radiotelephony communication.

9. The ICAO Rating Scale

One part of the puzzle is at the heart of it all: the ICAO Rating Scale itself.

To be clear, ICAO did a great job in creating a language standard for aeronautical communications. As a result, there is far more awareness of the role of language in the industry and much progress has been made. However, there are a host of well-documented issues with the ICAO Rating Scale which influence aviation language testing and assessment practice. To name three:

- It describes language which is undesirable on the radio, for example, grammatical complexity, idiomatic vocabulary and speaking at length. This results in test takers being rewarded for language use we do not wish to see.
- It does not adequately describe the features of language use we do wish to see on the radio, for example, precision, brevity and accommodation.

- It lacks explicitness. For example, what does ‘transition from rehearsed or formulaic speech to spontaneous interaction’ mean in the context of air-ground communication? Surely ‘transition from standard phraseology to plain language’ would be clearer?

These issues with the scale, among others, have contributed to widespread misinterpretation of the target of the ICAO LPRs which manifests in test instruments and assessment practice which are wide of the mark.

The ICAO Rating Scale is not necessarily a barrier to good testing, but there is room for improvement. Until we have an empirically validated scale which is representative of the target construct or, at the very least, clear ICAO guidance on how to capture the construct in aviation language test design, pilots and controllers will continue taking poorly-constructed tests that fail to address aeronautical radiotelephony communication.

10. The ICAO Rated Speech Samples Training Aid

There is a lack of appropriate models for aviation English test and task design.

After Doc. 9835, the most authoritative reference for aviation language assessment is the *ICAO Rated Speech Samples Training Aid* (RSSTA).[12] The RSSTA has played an important role in helping the international aeronautical community understand what spoken performances at various levels sound like, but it has had an undesirable side-effect: it has presented models for aviation English testing that have led to confusion around what we aim to measure and how to measure it.

A key objective of the RSSTA is to present samples of speech that are as representative as possible of test taker first language, professional background and language level. To achieve this objective, samples were gathered from far and wide with the caveat that ‘inclusion of a speech sample should in no way be interpreted as a judgement of the quality of the test tasks’. As with any project, the fundamental parameter for success is the quality of the raw material available and with the RSSTA, the result is a mixed bag. Let’s briefly explore three ways the RSSTA has influenced aviation English test design.

To make valid inferences about the ability of a test taker to communicate effectively on the radio:

1. We need test tasks that elicit performances which are:

- Representative of aeronautical radiotelephony communication; and
- Specific to the test-taker role.

[12] See: https://cfapps.icao.int/RSSTA/

Generic interview, picture description and discussion tasks may add value, but alone they are insufficient. Such tasks feature widely in the RSSTA, and this may explain, in part, the widespread misunderstanding of the construct we aim to measure with a corresponding failure of many tests to address the target of the ICAO LPRs.

2. We need tasks which specifically address listening comprehension in the context of aeronautical radiotelephony communication. Because:

- The RSSTA is a training aid for the assessment of speech (the clue is in the name!); and
- Assessment of listening through performance in speaking tasks is problematic;

The RSSTA does not provide appropriate models for listening comprehension tasks. This may explain, in part, why so many aviation English tests are under-representative of listening comprehension.

3. We need well-designed tasks administered by competent interlocutors. While the RSSTA features some good task design and delivery models, there are some poorer ones too. This may explain, in part, the poor standard of test design and administration which is so common in the field today.

Until we have clear, well-defined and appropriate models for aviation English test tasks alongside clear ICAO guidance on test design, pilots and controllers will continue to take poorly-constructed language tests that fail to address aeronautical radiotelephony communication.

For excellent guidance on aviation English test design, see the ICAEA *Test Design Guidelines*.[13]

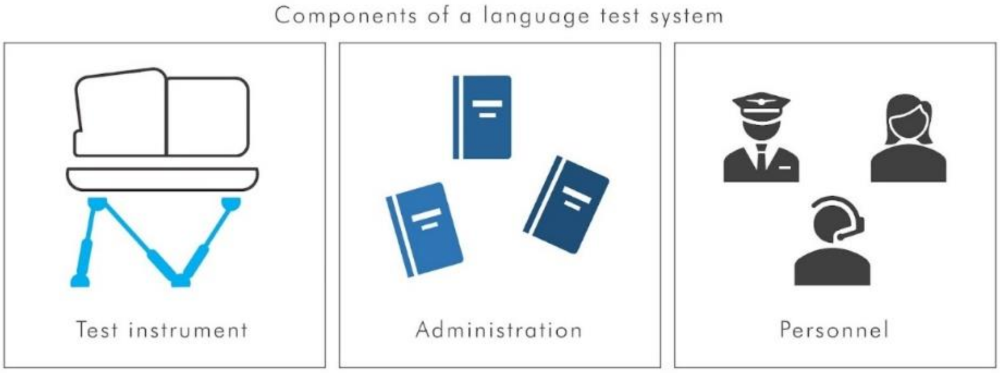
11. Untrained test personnel

Regulators give the job of language assessment to untrained, unqualified personnel.

Three things are necessary for aviation language testing to be effective:

1. A well-constructed test instrument comprising a range of tasks which together address domain-specific listening and speaking for the target population;
2. A system of administration including standards and procedures for test maintenance, personnel training and monitoring and test instrument validation; and
3. A team of trained, qualified personnel (managers, administrators, validation officers, interlocutors and raters).

[13] See: https://www.icaea.aero/projects/icao-lpr-tdg/guidelines/



If one of the above is faulty or missing, measurement will not be effective.

In the UK[14] and the USA[15] (and no doubt in other countries), Flight Examiners (FEs) and other personnel are permitted to conduct language proficiency assessments for licensing purposes. No test instrument is required. No training is required. No monitoring to ensure validity and reliability is required. Layperson assessments are made - quite literally - on the fly.

To be clear, aviation professionals are essential in aviation language test design and administration. But proficiency in English and an FE rating does not equate to language testing expertise. In an otherwise tightly regulated industry and in an assessment context where careers and safety are at stake, it is astonishing that regulators would permit untrained personnel to operate without monitoring and without a language test instrument.

Until regulators better understand language testing and ensure that reliable assessments are conducted by trained and competent personnel using valid test instruments, pilots and controllers will continue to take poorly-constructed language tests that fail to address aeronautical radiotelephony communication.

12. User perceptions and assessment literacy

There is a lack of assessment literacy coupled with competing priorities on the part of the organisations and individuals who use aviation language tests.

We would like airlines, ANSPs, ATOs and individual test takers to choose good testing practice over bad, but it is perhaps unreasonable to expect them to. To identify good practice, you need genuine interest and field-specific expertise. Test users are in the business of flying aeroplanes, controlling traffic, keeping costs to a minimum and getting on with their lives. Unless they are involved with test development and

[14] See: https://www.caa.co.uk/general-aviation/pilot-training-organisations/english-language-proficiency-testing-and-flight-crew-licensing/

[15] See: https://www.faa.gov/documentLibrary/media/Advisory_Circular/AC_60-28B.pdf

administration themselves, test users cannot be expected to be experts in language testing any more than I can be expected to be an expert in medicine when I visit the doctor.

Some argue that test users don't care. This is not true - many care very much. But we have to accept that the majority of test users are busy people with different priorities and motivations. I would go further and argue that any ignorance that we see manifesting in expediency and apathy has been strongly exacerbated by naïve and lazy regulation frequently exploited by amateur testing practice. Weak regulatory standards and sloppy testing have contributed to any perceptions held in the industry that language proficiency is not a central part of the safety picture, but a box to be ticked, a necessary evil.

Where does the onus to improve standards lie? Is it with ICAO who is responsible for the standard itself? Is it with the regulators who oversee it? Or is it the TSPs who, after all, are supposed to be the experts? Whoever is responsible, test users can only choose from what's available, if indeed they can choose at all. And where choice exists, poor practice lurks all too often behind edifices of reputability, and glossy websites with spurious claims for validity and promises of low prices and instant results which are so seductive that test users cannot see the wood for the trees.

Until regulators and TSPs better understand their professional responsibilities and make a more concerted effort to ensure the provision of relevant and meaningful language assessment that meets established standards for quality, pilots and controllers will continue to take poorly-constructed language tests that fail to address aeronautical radiotelephony communication. For the thousands of professional pilots and controllers who keep our skies safe, that is unjust.

Conclusion

This collection of short essays summarises my personal incredulity at the state of language testing in aviation. While revision of the standard and associated guidance would help, the real problem is a deep lack of assessment literacy amongst those responsible for the provision and oversight of testing services. The result is that many regulators fail to regulate which means that:

- Untrained individuals conduct casual layperson assessments on the fly;
- Non-commercial TSPs operate without adequate training and support; and
- Commercial TSPs (sometimes shamelessly) prioritise profit over professional responsibility.

The widespread amateurism is lamentable. Due to ignorance, wilful or otherwise, most aviation language testing does a disservice to not only the industry and individuals it aims to serve, but to aviation safety.

‘Language testing has evolved into an independent field that is characterised by well-articulated theories of validity and sophisticated validation methodologies’.[16] This translates into good practice in other domains (e.g. medical English, academic English), and it should in aviation, especially given the stakes involved. The fact is that it hasn’t: over the last 19 years, in spite of robust industry tools (e.g. ICAO recognition) and guidance (e.g. the ICAEA Test Design Guidelines) things are getting worse. Today, professional TSPs do exist, but they are vastly outnumbered by poor ones. The profoundly negative impact that this has on training continues to undermine the whole purpose of the ICAO LPRs.

Critiquing is easy. Putting things right is not. So, what to do? Perhaps the first step is raising awareness across all industry stakeholders - regulators, TSPs, industry organisations and test users - that something’s wrong. The more noise we make, the more likely it is that things will improve, and the more likely it is that any earnest efforts to promote better standards will gain traction. If you’ve taken the time to read any of the articles in this series, and if they have resonated with you in any way, then please don’t let your silence be interpreted as acceptance. The status quo is woeful, and things will only change if we call it out. Let’s hope that change comes before we have another Tenerife, Cali or Charkhi Dadri on our hands.

[16] Xi, X. (2014) *Methods of test validation* in Sohamy, E. & Hornberger, N. *Encyclopedia of Language and Education* (2nd Edition) Volume 7: Language Testing and Assessment (pp. 177-196) Springer: New York