

# AVIATION ENGLISH SERVICES

TEST INFORMATION HANDBOOK  
VERSION 1-2006



## Executive Summary

### BACKGROUND

- 1.1 Resolution A36-11 – *Proficiency in the English language used for radiotelephony communications* directs the Council to support Contracting States in their implementation of the language proficiency requirements by establishing globally harmonized language testing criteria.
- 1.2 While the first version of Doc 9835 - *Manual on the Implementation of ICAO Language Proficiency Requirements* published in September 2004 provided some guidance on testing, users of the manual including licensing authorities, air operators, air navigation service providers, language training and testing services have indicated that more detailed guidance on language testing was needed to effectively implement the language proficiency requirements.
- 1.3 Language testing is a complex and sophisticated professional activity requiring expert input at every stage: test development, implementation, administration, and scoring.

## SUMMARY REPORT

### LANGUAGE TESTING FOR COMPLIANCE with ICAO LANGUAGE STANDARDS: ISSUES AND CONSIDERATIONS

March 2006

PART A: Introduction.....	4
1 Statement of the Issue .....	4
2 Scope of the Report.....	4
3 Background: ICAO Standards in unregulated market .....	4
4 The Context: Test Development Standards .....	4
Part B: Test Characteristics.....	5
5 Test Purpose.....	5
• Diagnostic To identify strengths and weaknesses; assess gaps. ....	5
• Placement For placement into a tiered training program.....	5
• Progress To measure learning progress .....	5
• Achievement To measure overall learning .....	5
• Aptitude To assess ability to learn a new skill or knowledge set .....	5
• Proficiency To evaluate overall ability against a set of criteria.....	5
6 Test Characteristics.....	5
6.1 Language Focus .....	6
6.2 Delivery.....	6
6.3 Test tasks and Content .....	7
6.4 The Context II: ICAO Rating Scale and Holistic Descriptors.....	7
7 Assessing tests: Documenting Fairness .....	7
7.1 Validity .....	8
7.2 Reliability.....	8
7.3 Practicality .....	8
7.4 Test Washback .....	9
8 References.....	11
9 Appendix A: Checklist of Test Criteria .....	<b>Error! Bookmark not defined.</b>
10 Appendix B: ICAO Position Paper.....	<b>Error! Bookmark not defined.</b>

## **PART A: Introduction**

### **1 Statement of the Issue**

The International Civil Aviation Organization (ICAO) has adopted strengthened language proficiency requirements for flight crew and air traffic controllers operating along international air routes. As a result of the new Standards and Recommended Practices (SARPS), more stringent language testing requirements must be implemented by 2008, and pilots and controllers must demonstrate proficiency at the ICAO Operational Level 4 in order to maintain their license to operate internationally.

Organizations and individuals are seeking guidance on how to select, adapt, or develop appropriate aviation-specific English language proficiency tests which will ensure compliance with the ICAO Requirements.

### **2 Scope of the Report**

This report will briefly summarize the main characteristics of appropriate aviation-English tests and present the primary considerations in test selection and/or development.

Language testing is necessarily a rather complex issue, with professional expertise required at the level of selection and/or implementation. However, for the sake of convenience and ease, a very brief **checklist** for test evaluation is provided in Appendix A.

### **3 Background: ICAO Standards in unregulated market**

In March 2003, the International Civil Aviation Organization adopted Standards and Recommended Practices (SARPS) that strengthen language proficiency requirements for pilots and air traffic controllers operating along international air routes. The new language proficiency requirements clarify as a matter of an ICAO Standard that ICAO phraseologies should be used where possible, and that if ICAO phraseologies are not applicable then plain language proficiency is required. Codifying the use of plain language is a significant departure from both previous ICAO requirements and from de-facto practice. The ICAO language requirements establish minimum skill level requirements for language proficiency for flight crew and air traffic controllers in the use of both phraseologies and plain language. The minimum skill level requirements are embodied in the ICAO language proficiency rating scale and the holistic descriptors.

The new ICAO language SARPS create a significant testing and training requirement, particularly around the use of English. Reliable and valid aviation-specific English testing is not yet widespread or widely available, although more testing programs are coming into the market.

A complicating factor is that the language testing (and training) industry is both unregulated and professionally complex. In the high-stakes environment of aviation English testing, the lack of regulatory oversight is particularly problematic: the ICAO Standards and the 2008 compliance deadline create a market demand for testing services, but there is no regulatory body to provide oversight to the test providers, nor guidance to consumers.

### **4 The Context: Test Development Standards**

Language testing is a professional activity and is characterized by internally driven standards for test development, trialing, implementation, rating and reporting. There are currently few

organizations which provide test certification services, and no external industry requirements that a test undergo certification.

However, a number of resources guide test development, including the International Language Testing Association's (ILTA) Code of Ethics (included in ICAO Document 9835, Manual on the Implementation of ICAO Language Proficiency Requirements), the Association of Language Testers in Europe (ALTE) Principles of Good Practice, among others.

## Part B: Test Characteristics

### 5 Test Purpose

There are a number of different purposes for administering a test. Test 'purpose' influences the test development process. Some common language test types, related to test purpose, include the following, with brief descriptions:

- Diagnostic..... To identify strengths and weaknesses; assess gaps.
- Placement ..... For placement into a tiered training program
- Progress ..... To measure learning progress
- Achievement..... To measure overall learning
- Aptitude ..... To assess ability to learn a new skill or knowledge set
- **Proficiency ..... To evaluate overall ability** against a set of criteria

The ICAO Language Proficiency SARPS require *proficiency* testing. Proficiency testing is different from progress or achievement tests in that proficiency tests do not correspond directly to a learning curriculum. That is, it should not be possible to *directly* prepare, or study, (by memorizing information, for example) for a proficiency test. Proficiency tests require the test candidate to demonstrate their ability to do something, rather than simply measure how much of a quantifiable set of curriculum learning objectives

Proficiency testing is used to establish the competence of a candidate to exercise language skills in operational conditions. A working definition of a proficiency test, in our context, then, can be described as

- a set of structure events or procedures designed to elicit performances
- as samples of a candidate's language skills in a standardized way,
- to enable reliable inferences to be made concerning his or her level of competence,
- with the possibility of reproducing those skills at that level of competence consistently over time.<sup>i</sup>

### 6 Test Characteristics

After determining the testing *purpose*, there are a number of test development decisions which can be established, concerning *focus, delivery, method, task, and content*.

- Focus
  - **Speaking and Listening**
  - Reading or Writing
- Test Delivery Method
  - Direct
  - Semi-direct
- Test *Task*
  - Interview/Discussion
  - Role-play
  - Simulation
  - Questions/Answers
  - Discrete point items
- Test *Content*
  - Radiotelephony
  - Plain aviation language

## 6.1 Language Focus

Language proficiency testing evaluates a candidate's ability to use the language, to either speak it, understand it, write it, or read it. In the case of the ICAO Standards, candidates are required to demonstrate their speaking and listening proficiency.

## 6.2 Delivery

Speaking and listening proficiency can be assessed directly—through a direct interaction between the candidate and the assessor—or through semi-direct testing, in which test questions, or prompts, are pre-recorded and candidates record their responses individually, via a simple recording or computer-assisted.

Research shows that both direct and semi-direct test methods produce reliable results. However, the ICAO SARPS require the assessment of Interactions, which until now seems to require live, direct candidate-to-tester interaction.

Outside of the issue of Interactions, each test delivery method has a unique set of advantages and disadvantages.

	<b>Advantages</b>	<b>Disadvantages</b>
<b>Direct</b>	<ul style="list-style-type: none"> <li>• Ease of development</li> <li>• Provides direct interactions</li> </ul>	<ul style="list-style-type: none"> <li>• Difficult to administer</li> </ul>
<b>Semi-direct</b>	<ul style="list-style-type: none"> <li>• Easier to administer (able to administer to large numbers simultaneously)</li> </ul>	<ul style="list-style-type: none"> <li>• More difficult to develop</li> </ul>

### 6.3 Test tasks and Content

There are any number of test tasks or prompt types which can be used to elicit speech samples. In general, tasks which resemble real-life activities are most suitable. In the case of aviation English testing, however, the ICAO SARPS require not only proficiency in the use of the English (phraseology and plain language) used for radiotelephony communications, but also plain (aviation-related) English. It is important that any test elicit a range of speech sample, not limited to radiotelephony communication tasks.

### 6.4 The Context II: ICAO Rating Scale and Holistic Descriptors

Many features of language tests for the aviation industry are bound by constraints imposed by the ICAO Rating Scale and Holistic Descriptors. ‘

A description of some of those is found in Appendix B, an ICAO informational paper related to language testing. A summary of key points, related to test content is summarized here.

Radiotelephony communication involves the use of phraseologies and plain language. The language proficiency requirements are applicable to the use of phraseology and plain language. It is not the purpose of language proficiency tests to determine whether phraseology has been used accurately within an operational context; this is assessed during operational training and by operational examiners. Nevertheless, the holistic descriptors and rating scale do apply to the use of phraseology as well as plain language. Therefore, phraseology can be included in the range of stimulus in language proficiency tests, as long as it is only aimed at assessing language proficiency of the test taker. For example, phraseology can be used as a warm up or ice-breaking part of the test or as part of a script that will require the test taker to use plain language.

The ICAO Position Paper on Language Testing (2005) makes clear that ‘tests should provide candidate test-takers with sufficient and varied opportunities to use plain language in aviation work-related contexts in order to demonstrate their ability with respect to each descriptor in the Language Proficiency Rating Scale and the Holistic Descriptors’ (Attachment B).

A language proficiency test based only on phraseology is not considered valid because not all holistic descriptors and components of the rating scale can be assessed such as interactions, structure, vocabulary, etc.

## 7 Assessing tests: Documenting Fairness

The overriding concern of high-stakes test developers must be fairness. In language testing, fairness is interpreted in terms of validity—that a test indeed tests what is it supposed to test—and reliability—that the test gives consistent and fair results. Two other important traits include practicality and test washback. All tests must be evaluated in terms of their effect on test validity, test reliability, practicality, and washback effects.

## 7.1 Validity

Validity is a fundamentally important test characteristic, and it, basically, involves providing evidence to support the inferences that are made about an individual's English language proficiency based on their performance on a test. While validity can be thought of in overall terms, testers frequently examine the validity of a test in a number of different types of validity: such as content validity, construct validity, and concurrent or predictive validity.

Good testing practice requires, among other requirements, that a description of the validation processes used in the test development process be published as part of the documents relating to a test service (ALTE Principles of Good Practice.)

## 7.2 Reliability

Reliability refers to the stability of a test and test results; that is, evidence that the test can be relied upon to produce consistent results across different test takers in similar situations. There are a number of standard measures used in language test development to achieve this, including comparing two halves of a test to one another, or to compare the results on the test with the results by the same cohort of test-takers on another established test, among other methods.

### **Rater Reliability**

In speaking tests, an important aspect of reliability is rater reliability, and it is especially important to ensure inter- and intra-rater reliability. This is accomplished through rater reliability training and retraining, as well as sampling the ratings periodically to measure against the results of mentor expert raters. Reliability measures, again, are ensured through thorough test development, planning, and administration processes.

## 7.3 Practicality

Issues of test practicality impact test design in two ways: in terms of constraints imposed on the development process by available resources (funding, time, talent) and the practical aspects of implementing and administering the test into an established system.

### **Test Development**

Every test development project will face a unique set of certain constraints on the process. Issues like urgency, funding, resources, time, etc., necessarily impact the test development process. The commitment to test fairness, and validity and reliability, must be balanced against available resources and constraints.

A practical test is one that “does not place an unreasonable demand on available resources.” (ALTE 2001) If the resources are not available to support the development of a test with adequate attention to principles of good test design, then either (a) the test should be modified, or (b) the test administrators must make the case for an increase in resources or funding. When the latter option is not possible, then the test design should be modified to match the available resources while maintaining the high standards for validity and reliability evidence.

The ALTE Principles of Good Practice (2001) best summarize in detail how the issue of test practicality must be managed.

### **Test Administration**

A test may be valid, fair, and reliable, but if it is not also practical, then it is not usable or sustainable. The practicality of test administration must take into account particular national or local constraints. A three or four hour aviation English test may be reliable and valid, but it may not be practical in many instances.

## **7.4 Test Washback**

A final consideration of test usefulness concerns the ‘washback’ effect on training; that is, what effect on training ‘washes back’ from a test implementation?

Learners naturally want to be able to prepare for a test. If learners perceive that certain types of learning or practice activities will prepare them for a test, they will direct their energies to that, sometimes at the expense of activities which can actually help improve their language proficiency.

An example can be found in the (older forms of the) TOEFL test. The TOEFL included a large number of discrete-point (multiple choice, or error recognition) grammar questions, an indirect test method, if your goal is proficiency testing. As a result, students often did not perceive that communicative teaching methods would correlate to improved performance on the TOEFL, and rather preferred to spend time practicing TOEFL-like test questions. However, research showed that such activities did not correspond well to improve proficiency levels, a case of negative test washback.

In the aviation arena, an example may be found in an aviation English test which focuses too heavily on the use of phraseology or radiotelephony communications, at the exclusion of plain aviation language. In that case, learners may constrain themselves to focusing on memorizing more ICAO phraseology rather than on communicative language learning activities which will actually improve their English language proficiency, albeit in an aviation context.

## **8 Computer-assisted language testing**

A final important aspect of modern test regards the use of computers in the development, administration, and even rating of language proficiency tests. There are a number of ways that computers can facilitate language testing.

### **Administration**

Computers are very useful in the implementation and administration of language tests, allowing for tests to better replicate the ‘real world’ through simulation and role-play, for example. Additionally, computers permits larger-scale simultaneous administration of tests, in some cases.

## **9 Conclusions**

It is recognized that high-stakes language testing is complex. The best recommendation for organizations seeking to select, develop, or implement high-stakes aviation English language testing is to seek the input and advice of language testing professionals. Such support may be found within the Linguistic Departments of major universities. Additional direction is found in the ICAO Guidance Manual, Document 9835, in Chapter 4, which provides a chart of tester qualifications.

The Appendices to this Report provide more succinct review of the information necessary for appropriate test evaluation.

Aviation English Services is pleased to have been able to provide support to the organization and will be delighted to provide any other future assistance.

//end report

## 10 References

ACTFL. [www.actfl.org](http://www.actfl.org).

Alderson, J. C. Clapham, & D. Wall, (2001) '*Language Test Construction and Evaluation*'  
Cambridge: CUP.

*ALTE Principles of Good Practice for ALTE Examinations.*

[http://www.alte.org/quality\\_assurance/code/good\\_practice.pdf](http://www.alte.org/quality_assurance/code/good_practice.pdf)

*An Overview of the ACTFL Proficiency Interviews.* JALT Testing and Evaluation SIG  
Newsletter. Vol 1. No. 2. Sep. 1997, (p. 3 – 9). [www.jalt.org/test/yof\\_2.htm](http://www.jalt.org/test/yof_2.htm)

Davidson, Fred, and Brian K. Lynch. *Testcraft: A Teacher's Guide to Writing and Using  
Language Test Specifications.* Yale University Press. 2002.

Douglas, Dan. *Assessing Language for Specific Purposes.* Cambridge UP. 2000.

ICAO Document 9835: *Manual on the Implementation of the ICAO Language Proficiency  
Requirements.* 2001.

ICAO Position Paper: *ICAO Policy on Language Proficiency Testing.* (attached)

O'Loughlin, Kieran. *Studies in Language Testing: The Equivalence of Direct and Semi-  
direct Speaking Tests.* Cambridge UP. 2001.

---

<sup>i</sup> Mell, Jeremy. Paper presented at ICAO Regional Seminar on Aviation Language;  
Buenos Aires, Argentina, September 2005.