Report 01 – Description of Test Purpose, Specifications & Construction

Introduction

This report will describe the process by which the TEA Test Development Team (TDT) established the **test purpose** and **test specifications**. Additionally, it will consider how the **test method** and **test content** originated in respect of generating a fixed format that both reflected, and was appropriate to, the purpose and specifications.

Other reports will describe in more detail aspects of task design, item trialling, assessment, test administration, and test security.

Driving Forces

A number of initial 'driving forces' helped to form many of the critical aspects of initial test design. These are outlined below alongside a description of their implications for the construction of TEA. It was agreed that test development should adhere to the guidance from ICAO to be considered valid on a conceptual level.

References to the *Manual on the Implementation of ICAO Language Proficiency Requirements* here are to the 2nd Edition (of 2010) and will simply be '*9835*'.

Driving Force	Implication for the TEA
1 **ICAO's description of language proficiency** In attempting to define what the LPRs are designed to measure, 9835 declares that language proficiency can be seen as a combination of abilities. It states that 'communicative competence' is the minimum language requirement for operational personnel. This includes aspects of linguistic (productive & receptive), sociolinguistic and pragmatic competences, and a recognition that errors may be present in performance, but need not interfere with successful communication. Communicative competences can be measured through language performance.	

9835 states:

2.3.3.1: "All the competences needed for language proficiency are "constructs" of mental and physical abilities and they are not directly observable. They can be inferred in individuals only by observing the language performance of those individuals. In performance, other factors may impact language proficiency, for example, levels of attention, mood, stress, verbal working memory and verbal processing abilities. These factors will, in turn, influence levels of performance in the areas of fluency, comprehension and interaction."

2.3.3.2.: "Performance then is not the same as competence, but provides the only opportunity by which competence and language proficiency can be inferred and assessed…."

6.3.2.5.: "The more directly a test performance is related to target performance, the more a test can be considered a proficiency test. For example, test administrators interested in an individual's speaking skills should arrange for an assessment of that individual's performance on a speaking task. Using this approach, speaking skills may be directly assessed during an interview or conversation or role-play, or are based on a recorded sample of actual speech."

Given this theoretical framework, the TDT needed to define which competences were to be measured by the test for it to be considered appropriate (valid) to the context.

The test designers wanted to ensure, as far as possible, that the test measured only language proficiency, and not intelligence, logical thinking, operational knowledge or any other construct which would unfairly affect the assessment.

Holistic Descriptor 'd' emphasises the need for strategic competence to overcome linguistic problems and, in 9835, it is stated that "appropriate responses must be delivered efficiently and a rapid response time is expected. The interactions skill refers to this ability, as well as to the ability to initiate exchanges and to identify and clear up misunderstandings." (4.6.7.)

In recognising the requirement for the testing of

interactive competence, the TDT had to consider the most appropriate method of test-delivery, and focus on communicative, interactive tasks that elicited language that allowed for the assessment of the candidates' ability to manage the interaction.

2
Standard Phraseology & Operational Language versus plain English

6.2.8.4.: "Radiotelephony communications require not only the use of ICAO standardized phraseology, but also the use of plain language. Phraseology is the formulaic code made up of specific words that in the context of aviation operations have a precise and singular operational significance. Plain language is defined in ICAO documents as "the spontaneous, creative and non-coded use of a given natural language.""

6.3.2.8.: "ICAO language provisions require proficiency in the use of standardized phraseology and in the use of plain language. The assessment of standardized phraseology is an operational activity, not a language proficiency assessment activity. While an aviation language test may include phraseology to introduce a discussion topic or make interaction meaningful to the test-taker, it is important that tests elicit a broad range of plain language and not be limited to tasks that require standardized phraseology. The focus of a language proficiency test for compliance with ICAO requirements should be on plain language."

The target language to be elicited is plain English in an aviation context.

The ICAO Rating Scale was clearly designed to assess plain English, rather than phraseology (for example by measuring paraphrase, idioms, and register).

6.3.2.9.: "An aviation language proficiency test has different aims than a phraseology test. It is acceptable that a test contains a scripted task in which phraseology is included in a prompt, but the test should not be designed to assess

phraseology."

6.3.2.10 "The test should not be designed to evaluate the technical knowledge of operations. Language tests should not assess either operational skills or the specific technical knowledge of operations. A language test is not an operational or technical knowledge test."

Although every attempt should be made to avoid eliciting phraseology, it may be used in prompts within a task.

The TDT recognised that ICAO standardized phraseology and technical aviation language is
a) a pre-requisite for pilots and controllers working internationally, and
b) assessed during their training and licensing.

It is, therefore, not an aspect of aviation communications to be assessed in this testing context.

Dr. Marjo Mitsutomi served on the ICAO PRICESG (Proficiency Requirements in Common English Study Group). In her 2001 collaboration with O'Brien, Mitsutomi describes aviation language proficiency as comprising of 3 parts:
1. ATC phraseology
2. English for Specific Purposes (operational and technical language specific to pilot-controller communications)
3. English for General Purposes.

Of English for General Purposes, the authors state that:

"The ability to communicate when there is no prescribed script (ATC phraseology) is critical to safety. In practice this means that pilots and air traffic controllers must have the ability to achieve mutual understanding through the use of their general language ability to get their messages heard and understood…. this ability to negotiate meaning at all times is the key to communicative competence." (p.13)

"Recognizing then, even if reluctantly, the need for competency in general English to complement the use of ATC has been the driving force behind the ICAO PRICE (Proficiency Requirements in Common English) Study Group and the FAA PEC (Pilot English Competency) Working Group for the last few years. The task of these groups has been to define the minimum level of proficiency in English needed to communicate safely at all times. ICAO has pioneered the way by already describing this minimum level of proficiency that facilitates speaking and understanding English in usual and unusual aviation-related contexts." (p.14)

It was clear that Mitsutomi & O'Brien were referring to the need for the assessment of plain English.

3
Work-related Context

6.3.2.8: "The idea of a work-related context can be interpreted in different ways…. The narrow view would seek to replicate radiotelephony communications including both phraseology and plain language, as closely as possible. The broad view would elicit samples of interaction and comprehension on those topics occurring in radiotelephony communications without resorting to replicating radiotelephony communications. These could be of a general piloting and controlling nature and involve question and answer routines, short reports or problem-solving exchanges, or briefings and reports."

Although both narrow and broad views of 'work-related context' could be considered appropriate to the testing situation, the narrower the approach, the narrower the language elicited –i.e. a narrow approach was more likely to elicit operational and procedural language that may, at least partly, be non-assessable using the ICAO Rating Scale. Thus, in terms of validity, a broader approach would be more likely to meet the purpose of general proficiency testing while avoiding the assessment of operational competence.

4.6.6.: "In air traffic control communications, pilots rely on the clear and accurate

information provided to them by controllers for safety. It is not sufficient for air traffic controllers to be able to handle most pilot communications; they must be ready for the unexpected. Similarly, pilots must be able to understand air traffic controller instructions, especially when these differ from what a pilot expects to hear. It is during complications in aviation that communications become most crucial, with a greater reliance upon plain language."

It was appropriate to disregard content which was wholly 'job-specific' (see below).

Furthermore, the TDT recognised that the type of language to be tested in this context was language that would, hopefully, never be needed. Although the content needed to be work-related, it would be artificial by nature, exposing candidates to lots of problem situations when, realistically, they might never have to deal with any emergency situations that require plain English during their entire career.

4
The ICAO Holistic Descriptors and Rating Scale

6.3.2.1.: "The test should be designed to assess speaking and listening proficiency in accordance with each component of the ICAO Language Proficiency Rating Scale and the Holistic Descriptors."

The imposed assessment criteria had the following implications for the TDT:

- The ICAO Rating Scale was clearly designed to assess plain English, rather than phraseology (by measuring, for example, paraphrase, idioms, and register).
- The Scale helped to form the language abilities

that should be measured (e.g. "Is able to comprehend a range of speech varieties (dialect and/or accent) or registers". Relevant grammatical structures and sentence patterns should be determined by language functions appropriate to the task.

- The test must elicit language assessable by the Scale. Therefore, the test had to:
    - Refer to familiar, common, concrete and work related topics
    - Refer to unfamiliar work-related topics
    - Expose candidates to a wide variety of international accents
    - Expose candidates to a linguistic or situational complication or unexpected turn of events
    - Allow candidates the opportunity to demonstrate discourse management strategies
    - Allow opportunities to display knowledge of idiomatic expressions, register & to speak at length.

    The TDT considered that were these aspects not included in the test, the language elicited could not be rated by the Scale.

- Rather than awarding a global performance score, the Scale is clearly designed for *profile marking* in which raters assess 6 aspects of linguistic ability based on interpretations of performance descriptions. Therefore, quantitative assessment through measurable, objective data would be difficult to both implement and justify (validate).

5
ICAO Language Functions & Domains

Appendix B of 9835 (Communicative Language Functions, Events, Domains and Tasks associated with Aviation) lists 116 functions associated with pilot-controller communication. The vast majority relate to both pilots and air traffic controllers.

The TDT would aim to elicit such functions within the test in order that test content reflected work-related content.

9835 lists 'events and domains' that characterise communications between pilots and controllers in routine & non-routine situations.

When item-writing, by focussing on the elicitation of key language functions associated with pilot-controller communication the test need not distinguish between pilot and controller candidates.

The TDT would be able to maintain focus on relevant, work-related test content by using these domains and events (topics) to form the content for task & item writing.

6

High Stakes

From 6.2.2.: "The results of language testing can have a serious impact on both individuals and organizations. A pilot or controller operating internationally who does not demonstrate compliance with the ICAO language proficiency requirements may be denied a licence to operate internationally, a consequence which may severely impact the career of that individual as well as the staffing requirements of the airline or air traffic service provider for whom the individual works.

"…there are economic factors to consider. State authorities, airlines and service providers have no funds to waste on inadequate or unproven tests, nor can they afford to lose otherwise competent staff as an outcome of inadequate testing. Ultimately, they cannot afford accidents attributable to ineffective pilot/controller communication.

"…it is vital that language testing for licensing purposes comply with best practices and address the specific requirements of aviation operations."

The overriding concern of test designers was to help promote aviation safety through the construction of a valid, reliable testing system. However, *practicality* is also a major consideration for test developers. The key notion was one adopted by ICAO – "Affordable Safety". Stakeholders would need to invest heavily in appropriate training programs in order to help their personnel achieve Level 4, 5 and 6. It was critical then that the test offered an affordable, practical means of assessing personnel, in order that airlines and ANSPs could maximize their training resources.

6.3.2.8.: "A further step toward providing test-takers with a familiar aviation-related context would be to customize the tests for controllers or pilots. Thus, controllers would have the possibility of taking tests using or referring to a tower, approach or

en-route environment; similarly, pilots would be able to take tests using or referring to an approach procedure. These should be seen as adaptations in the interest of the comfort of the test-taker, not as specialized tests of distinct varieties of language proficiency."

6.3.5.9.: "In the case of semi-direct test prompts (which are pre-scripted and pre-recorded), there should be adequate versions to meet the needs of the population to be tested with respect to its size and diversity. Tests with specific pre-recorded or pre-scripted questions or prompts require multiple versions."

6.3.3.3.: "Test washback refers to the effect a test has on a training programme or on students' behaviour. TSPs should demonstrate that their test will have a positive effect on training and that their test will not encourage training that focuses on memorization and test preparation rather than on building proficiency..... Test-takers naturally will want to prepare for a test. While aviation language test-takers can memorize phraseology, they cannot acquire language proficiency as described in the ICAO LPRs simply by memorizing words and phrases. If pilots or controllers think that certain types of narrow learning or practice activities will best and most readily prepare them for a test, they will be inclined to direct their energies to such activities, potentially at the expense of activities that can genuinely improve their language proficiency.

With respect to the needs and wishes of the stakeholders, the following implications emerged from both 9835 and discussions with industry personnel (both decision-makers at CAAs and ANSPs, and pilots and controllers):

Delivery & Rating
While test delivery to 2 candidates simultaneously may provide opportunities for role-plays and interaction, the advantages are outweighed by the practical issues, the difficulty in candidates of different abilities

interacting, and the potential hierarchical issues (e.g. a senior Captain paired with a junior First Officer). A short test, delivered one-to-one would be more practical considering both the difficulty in stakeholders finding suitable time for test-taking and the potential consequences for individuals.

Furthermore, a test that could assess across the full range of performance criteria (levels 1 – 6) would be more appropriate to the test users.

Content
Tests aimed solely at commercial pilots or en-route controllers would not be appropriate for reasons of language (see Section 3 *Work-related Context* above) and practicality - producing a test specifically for en-route controllers would only be of use if the candidates remained in that position for the period their test scores were valid. Stakeholders did not want to consider extra testing as a consequence of re-licensing.

Versions and Applicability
The scale of testing requirements indicated that the test would be delivered in multiple locations to a large candidature over a number of years. Live test materials have a limited shelf life & need replacing regularly to maintain confidentiality of materials (and therefore, reliability of results). Therefore, it

was essential to be able to produce multiple versions of the test, and that the content be globally applicable i.e. not culturally-bound so as to disadvantage some candidates.

Tasks needed to be standardised to ensure all sets were of a similar level of difficulty and elicited similar language.

Administration & Security
With aviation safety and livelihoods at stake, there would be a risk of candidate collusion or cheating. This would be reduced by having multiple versions of the test (see above) & by creating a bank of materials that could quickly replace any versions of the test that had been compromised (for instance if they were stolen from an exam centre). It also influenced the method of test delivery (the TDT considered computer-based tests too insecure given the consequences of the test results.)

7
Test Method & Delivery Method

6.7.10.: "Direct, communicative proficiency tests of speaking and listening abilities are appropriate assessment tools for the aviation industry and will allow organizations to determine whether flight crews and air traffic controllers are able to meet the ICAO language proficiency Standards".

Direct human delivery was both acceptable and most appropriate to the high stakes nature

6.3.2.6.: "If comprehension is assessed through a specific listening section with individual items, it should not be done to the detriment of assessing interaction.

"Some language tests evaluate listening during an oral interaction such as a conversation, interview or role-play. Other language tests evaluate listening separately, in some cases via a series of individual listening items. An example of an individual listening item, in the aviation language context, might require a test-taker to listen to a pre-recorded conversation between ATC and a flight crew to identify relevant pieces of information."

9835 defines proficiency in Comprehension as "the ability to recognize and understand speech. Development of this skill will result in decreasing difficulty when dealing with complex discourse, with unexpected or unfamiliar topics, unfamiliar accents or delivery styles and with unfavourable conditions of reception (due to background noise, etc.). Proficiency in comprehension can be characterized by the degree of detail and speed of understanding." (p.29).

of the context, allowing a professional language examiner to fully explore the language proficiency of the candidate. Delivery and/or assessment by computer would incur many doubts about suitability, validity, reliability and security.

Face-to-face interviews that effectively elicit an appropriate sample of spoken language for assessment purposes through direct and semi-direct methods were appropriate.

The TDT considered *integrative testing* a method appropriate to the assessment of Comprehension and Interactions since it allowed candidates to demonstrate understanding through immediate oral production – an approach authentic to the target language situation (at work). Further,

'live' comprehension would allow for demonstration of discourse management strategies and proficiency across all six language profiles.

A separate listening test would demand the use of construct-irrelevant skills (such as reading, writing or keyboard skills) and would present challenges to test administration.

6.3.2.7.: "Proficiency tests that are administered directly may use face-to-face communication in some phases of the delivery but should include a component devoting time to voice-only interaction. Voice-only interaction is an important characteristic of aeronautical radiotelephony communications; when a pilot and a controller interact, they cannot see each other. Directly administered proficiency tests should simulate this condition of "voice only" in at least a portion of the test…. An appropriate strategy may be to incorporate both direct and semi-direct methods in a single testing system."

A semi-direct system of 'live' listening and interaction would allow for a voice only component in which the examiner remained silent while administering the semi-direct input.

8
Rating

6.3.4.1.: "Some speaking and listening tests rate performance during the test. Others record the test performance and rate performance later. Both rating methods are acceptable."

It was appropriate to consider the value of different rating and interlocution formats with either simultaneous or retrospective rating considered valid.

It would be necessary to record tests.

6.3.4.2.: ""Best practice in language proficiency assessment calls for at least two trained and calibrated raters, at least one of whom is a language expert… Using at least two raters reduces the possibility of rater error and helps to ensure a comprehensive evaluation of each test-taker…. Ideally, an aviation language test will have two primary raters — one language expert and one operational expert — and a third rater who can resolve differences between the two primary raters' opinions."

It would be essential to find and train raters with a language testing background. It would be crucial to have double-marking procedures in place.

The notion of 'two primary raters' would need to be considered carefully from a practical viewpoint, in terms of location, time and cost. A two-rater assessment could be compromised by unreliable factors such as status and experience: a power relation or other role conflict can lead to unreliable results.

In this context, training suitable operational personnel to become language raters would be appropriate. It would also be appropriate to include operational experts in standard setting.

6.3.4.3.: "When evaluating language proficiency tests, consistency in the rating process is critical. Unlike other forms of testing, in which one response to a question is correct and another response is incorrect, evaluating language proficiency relies upon subjective judgements by raters. In this context, consistency is achievable through training and experience but easy to lose without regular audits of raters and rating teams."

It would be necessary to train, certificate, standardise and re-certificate examiners to meet standards of reliability in assessment.

9
Administration & Security

6.3.5.8.: "Testing organizations should ensure that people do not have access to specific test content or questions before the test event. In addition, TSPs should ensure that test scores are kept confidential….The ongoing reliability, validity and confidentiality of a language proficiency testing system will depend heavily on the test security measures that are in place. Testing organizations should protect test-item databases and provide secure storage of scores and test materials. They should require, establish and maintain formal commitments to confidentiality and integrity from test developers, administrators, raters, information technology personnel and any other staff who are involved in any aspect of the testing process. Other necessary security measures during test administration should prevent:
a) communication between test-takers; b) communication between test-takers and people elsewhere during the test (for example, by use of a mobile telephone); c) impersonation of others; and d) the use of false identities."

It would be necessary to train test administrators thoroughly in aspects of security, data collection and record-keeping.

It would be necessary to develop technologies to allow for secure data collection and transfer.

It would be necessary to consider the security issues surrounding test certificates and their issuing.

Specifications

The following table briefly outlines some of the key components of TEA formed as a result of the considerations above.

Test Purpose	To effectively elicit language assessable by the ICAO Rating Scale (levels 1 – 6) in order to assess the plain English language proficiency of both pilots and air traffic controllers working in an international context
Test Users (Candidates)	Pilots and air traffic controllers operating in an international context
Test Focus	Speaking, understanding and interacting in plain English
Test Length	Approximately 20 minutes
Level of Difficulty	From levels 1 – 6 as laid down by ICAO's Rating Scale
Test Content	Based on the Language Functions and Domains described in 9835
Test Context	Common, concrete and work-related topics and broader aviation topics
Task Types	Communicative, interactive tasks
Test Delivery	One-to-one: direct (face-to-face with interlocutor) and semi-direct (recordings of international speakers on CD)

Examiners	Trained & certified to rate and/or deliver tests Language proficient Suitable knowledge of aviation terminology Jointly responsible with test administrators for test security, including candidate identification
Rating	Conducted either simultaneously by examiner or later via the test recording Double- and triple- marking conducted via the test recording
Administration	Responsible for scheduling and candidate application, and test-day organisation (including taking candidate photographs) Responsible for secure storage of test materials, collection of data entry and file transfer Jointly responsible with examiners for test security including candidate identification
Security	Secure online database, and audio transfer system Certificates only printed centrally containing candidate's photograph (taken immediately prior to the test) and biographical data

Language competences

The TDT needed to define which competences were to be measured by the test for it to be considered appropriate (valid) to the context. 9835, was prescriptive in defining which competences should be measured[*]:

Productive competences

- Talk about familiar, common, concrete and work-related topics specific to the candidate's role in aviation

- Talk about familiar, common, concrete and work-related topics common to pilot – controller roles in general

- Talk about aviation-related topics in a broader context

- Use a range of basic and complex grammatical structures as appropriate to the function of the task

- Use a range of work-related vocabulary

- When lacking vocabulary, use circumlocution strategies

- Produce connected stretches of language, sometimes at length

- Use a range of phonological features (sound, stress, rhythm, and intonation) to produce speech intelligible to the international aviation community

Receptive competences

- Understand the specific details of short messages delivered by both pilots and controllers in plain English in non-routine situations at different phases of flight (tower, ground, departure, en-route, approach)

- Understand a range of native and non-native speakers in terms of accent and rate of speech

- Process linguistic difficulties such as

    o tense
    o modality
    o lower frequency work-related vocabulary
    o negation
    o contraction

- Recognise the illocutionary force (the communication purpose) of the speaker

- Understand and respond to short messages describing linguistic or situational complications or an unexpected turn of events (in an aviation context).

[These 'macro-abilities' are assessed by targeting sub-skills such as distinguishing word boundaries, identifying stressed and weak forms, and recognising reduced forms. It was also essential to consider which other potentially significant 'abilities' may already be implicit in previous aviation phraseology training (such as the ability to recognise elliptical forms of grammatical units and sentences), and which abilities may be inappropriate to test. The assessment of inference skills is one of the biggest challenges to testers, but items that go beyond literal meanings are seen as valuable in a comprehensive assessment of listening. In terms of defining listening constructs for this context, however, abilities such as inference or deduction, from either operational, world or contextual knowledge, do not seem appropriate to the field of aviation in which precision in comprehending short, transactional, radiotelephony messages is essential. The further complication in measuring inference skills in plain English for Aviation is that, by encouraging interpretation of a wider communicative context, we may encourage procedural or operational responses that move away from the assessable language domain.]

Interactive competences

- Where necessary, demonstrate discourse management strategies to resolve misunderstanding

- Manage and maintain the speaker-listener relationship

- Respond immediately, appropriately and informatively

While acknowledging the guidance of 9835, the TDT considered the testing of both 'sensitivity to verbal and non-verbal cues' and comprehension of "cultural subtleties" beyond the scope of a formal, standardised language test.

*** It was assumed that candidates possess the required language proficiency in Standard Phraseology appropriate to aviation radiotelephony communications.*

Rationale for Test Method of the TEA

The TDT considered the oral proficiency interview as appropriate to the context because it:

- Is a face-to-face interview that effectively elicits an appropriate sample of spoken language for assessment purposes through direct and semi-direct methods.
- Attempts to focus on functional speaking ability and communicative competence in English regardless of specific subject 'knowledge'.
- Tests general ability for future use.
- Is objective and non-culturally specific.
- Is a cost-effective and time-efficient means of testing language proficiency.

Construct Rationale Statements

1) Communicative testing must be tests of performance rather than knowledge – we should be more concerned about whether candidates can use the language effectively than how much they know.

2) Parts of tests in this context should focus on common, work-related topics familiar to the candidate (i.e. tasks specific to aviation role); parts should focus on the unfamiliar and unexpected. Therefore, both a narrow and broad view of 'work-related' is appropriate.

3) The target language use situation is ill-defined since the target for elicitation is 'plain English in an aviation context', and not phraseology/operational language. It is therefore better to avoid attempting to elicit language truly authentic to the target language domain (i.e. a combination of standardised phraseology and plain English) but develop tasks and items which elicit the language abilities (competences) appropriate to that domain.

4) Although potentially considered 'purer' tests of comprehension, separate listening tests (which measure listening comprehension only) are not as appropriate to this testing situation as integrative tests which allow for immediate interaction through oral production. The focus is on language use rather than language knowledge (with an emphasis on assessing the processing of language as opposed to assessing knowledge about elements of language).

5) Collaborative listening tasks in interactive, interview-type tasks demand actual communication between two participants and are therefore appropriate in this context.

6) In this context, it is appropriate to focus on a short-text processing approach (to mirror pilot-controller communications). The recognition and processing of clearly-stated details is most crucial. In a context of 'non-routine' and 'unexpected complications', it is not relevant to

ask candidates to relate the linguistic information to a wider context or process inferential meanings since the concrete details of the situation are what demand comprehension.

7) Listening texts can challenge phonological short-term memory but this is considered a construct of proficiency in listening comprehension (a variable in language proficiency), much like aptitude or motivation, rather than an individual trait independent of language ability. In non-routine or emergency situations, pilots and controllers would need to rely on their phonological short-term memory to process messages.

8) Exposing candidates to a range of speech varieties demands a semi-direct approach.

9) Test tasks should attempt to include anything which is dependent on linguistic knowledge and exclude anything which is dependent on general cognitive ability, operational knowledge, or irrelevant skills (such as reading, writing, or computer skills).

References

Mitsutomi, M. & K. O'Brien (2003) *The Critical Components of Aviation English* University of Redlands

ICAO Document 9835, 2nd Edition. (2010). *Manual on the Implementation of ICAO Language Proficiency Requirements*