Report 02 – Overview of Expert Judgements & Action Taken in Test Development

TEST OF ENGLISH FOR AVIATION

Introduction

In every stage of language test development, it is necessary to rely on human judgements. Clearly, so-called 'expert' judgements are desirable in analysing and defining test purpose and specifications, in task and item creation and revision, and in classifying and describing performance criteria. Although data-gathering through multiple trialling and administration is a crucial aspect of decision-making, it is essential that test developers have access to relevant, experienced *opinions* to allow the process to begin and be refined, stage after stage. Alderson (1993) goes further, describing the need for professional judgements as continuous, even in tests which purport to be objective assessments of language:

> "*language testing is an area of applied linguistics that requires judgements at every level of activity and every stage in test development and validation. Testers have to judge whether test specifications are fit for their purpose, whether test content reflects the test's specifications, whether the test method is appropriate for the test's purpose, whether scoring criteria are appropriate, and whether candidates' performances meet those criteria*." (p.46)

A reliance on human judgements quite naturally raises two questions to consider. Firstly, given the inevitability of judgements affecting crucial aspects of test design and operation, it would seem vital to attempt to ensure some degree of consistency. However, it appears that little research has been conducted into this area. In his 1993 paper *Judgements in Language Testing*, Alderson demonstrated that variability between expert opinions could seriously undermine test validity. Alderson, Clapham and Wall (1995) noted that "quite often experts do not agree with each other" (p.175), leading to a potential array of avenues to be explored beyond initial expectations. Secondly, the notion of 'expert' must be scrutinised further, since one can never be satisfied with the term without relevant definition to the specific context. Hughes (1989) identifies experts as "people who are familiar with language teaching and testing but who are not directly concerned with the production of the test in question" (p.27). Clearly, in the context of testing Plain English for Aviation, there are a number of sources of potential 'expertise' – but experts in operational language might well have little or no knowledge of language proficiency testing, and vice versa.

Principle 5 of ILTA's Code of Ethics infers that language testers should "continue to develop their professional knowledge", specifying that they should take opportunities to "interact with colleagues and other relevant language professionals as an important means of developing their professional knowledge". Continued learning and development is fundamental to the test development team's (TDT) professional role and failure to seek improvement would be a disservice to TEA candidates. Since high-stakes international language tests like TEA can impact significantly on individuals and organisations, the TDT must seek to involve as many stakeholders as possible, and be prepared to review and take into account how they use the test and what they think about it.

Across every stage of test development, the TDT have engaged with a variety of aviation and linguistic professionals. Through surveys, focus groups and general interaction with decision-making stakeholders, operational personnel, English language experts and language testing experts,

opinions and comments have been fed back into the development process. In some cases, the feedback has proven instrumental in action being taken; in others, it has provided a useful catalyst for internal critical analysis and reaffirmation of the testing procedures and processes of TEA.

Data Collection

Qualitative research is often used for policy evaluation since it can answer certain important questions more effectively than quantitative approaches. Principle methods of qualitative research are the collection of feedback through the use of surveys, interviews, and 'think aloud' verbal protocols**.** The results can help decision-makers to understand how and why certain outcomes were achieved and reveal important answers about relevance, unintended effects and impact of the test. On the basis of the data collected, it is possible to draw consistent and more reliable conclusions about the judgements that have been made in the development process. In comparing data, the strength of the judgements made can be assessed, potentially leading to a refinement of methods.

The table below catalogues a selection of critical responses to a variety of qualitative analyses conducted with operational personnel during test development, including candidates (the subject matter experts or SME), the stakeholders at decision-making level for CAAs and ANSPs (test users or TU), English language experts (ELE), Language Testing experts (LTE), and TEA Examiners (TEX). They are expressed as sentiments rather than exact quotes.

Expert Criticism	Test Development Team (TDT) Response
About TEA generally…	
SME: *Pilots do not have to answer such questions in their job…. The language demanded in the test is not the language used in the target language situation.*	The question of authenticity has long been a complicated one in language testing. Fulcher & Davidson tackle the issue in Language Testing and Assessment (2010): "It has sometimes been said that tests are in some way better or more valid just because they are direct. But this is not the case, for, as Bachman (1990:287) has pointed out, directness is problematic in language testing, as 'language is both the object and the instrument of our measurement'….. arguments over task authenticity that dominated the late 1970s and 1980s are no longer meaningful for us. Proponents of the communicative approach to language testing argued that only tasks that mirrored language use in the 'real world' should be used in communicative language tests, reflecting the actual purposes of real-world communication,

	in clearly-defined contexts, using input and prompts that had not been adapted for use with second-language speakers… At the time, Alderson (1981) argued that this was a 'sterile argument', and we have since realized that authenticity, even conceived of as matching test method facets to facets of similar tasks in the real world (Bachman and Palmer, 1996), does not make tasks automatically valid through directness; it means only that we may be able to model test-taker behaviour in ways that allow us to observe the use of processes that would be used in real-world language use… we need to look at how items are used within specific tests in order to judge whether responses to those items (and the resulting score) would support the inference from the evidence to the claims we wish to make about the underlying knowledge or ability of the test-taker." (p.63) Thus, in order to study 'evidence' from candidate responses to test items and assess whether it supports the claims TEA is making about candidate abilities, various studies were conducted such as, for example, the use of Observational Checklists of test transcripts to measure language functions elicited during the test.
SME: *It would be better to include role-plays with operational examiners.*	In the early stages of test development, role plays were trialled. The obvious advantages of using role-plays (relevant communicative language task, easy to produce multiple versions) were hugely outweighed by the disadvantages: • As the 'roles' were narrow and specific, the language elicited was more memorised and procedural than spontaneous plain English. • It was difficult to deliver instructions to the candidate without a) it potentially being a reading comprehension test, and b) giving the candidate the language they needed to complete the task itself. • The role of the interlocutor was impossible to standardise since role-plays can proceed in multiple directions. • Candidate reluctance to participate in an imaginary 'game' – this factor seemed particularly culture and individual-dependent.

LTE: *The content validity of the test is compromised by the attempt to measure proficiency across such a broad range of abilities (in this case ICAO Levels 1 – 6). There should be a separate test to measure proficiency at each level.*	The TDT recognise the potential weaknesses of measuring such a broad range of language proficiency in one test. However, the TDT maintain that it is possible (as evidenced by the range of language elicited by TEA candidates given the same test – see ***Report 06 – Content Analysis: Language Functions & Language Elicited in TEA***) and the industry demands it for reasons of practicality and cost.
LTE: *The content validity of the test would be higher if you tested approach controllers on matters solely concerning the role of approach controllers, as with tower and en-route controllers, private pilots, etc, etc…*	A similar criticism was made about commercial & private pilots – should PPL-holders be subjected to items related to commercial aviation issues? As above, the content validity may have been compromised by the decision to test all pilots and controllers with one test but the following factors strongly influenced the TDT: • trialling showed that the narrower the test focus, the narrower the range of assessable language elicited – counter-productive to the main objective of eliciting a broad sample to be assessed using the ICAO Descriptors; • discussions with SMEs confirmed that private pilots might well need to understand radio communications between commercial flights and ATC; • discussions with industry stakeholders suggested that different tests for different positions/licences would be costly and impractical. In November 2011, in the Official Journal of The European Union, the following directives are published under 'FCL.055 Language proficiency' (page 12): *(d) Specific requirements for holders of an instrument rating (IR). Without prejudice to the paragraphs above, holders of an IR shall have demonstrated the ability to use the English language at a level that allows them to:* *(1) understand all the information relevant to the accomplishment of all phases of a flight, including flight preparation;*

	In order to react positively to the feedback, the following adjustments to item writing were considered, trialled and implemented: • role-specific questions on common, concrete, work-related topics • a balance of items to reflect commercial / non-commercial topics • a balance of recorded items to reflect different roles & different phases of flight.
SME: *The face validity of the test is very low. Why isn't there any phraseology in the recordings?*	In the initial stages of test development, the TDT agreed that a test that tried to combine the 'codes' of phraseology and plain English would not be appropriate for the following reasons: • *9835* clearly states that it is plain English in an aviation context that should be assessed. • Separating the two 'codes' appeared impossible. • The test should measure only language proficiency in plain English, rather than Standard Phraseology or operational knowledge (or intelligence, logical thinking, or other construct-irrelevant factors). The testing of Standard Phraseology needs to be assessed by operational experts using a different set of criteria (not the ICAO Language Proficiency Scale). • Although the test may appear inauthentic to operational personnel, the test content *could* be considered valid if the language functions and domains elicited were appropriate to the target-language context. As trialling began, it was clear that many scripted items which contained phraseology as a prompt, elicited only operational language or a combination of the two 'codes'. See ***Report 01 – Description of Test Purpose, Specifications & Construction*** for further information.

TU: *The comprehension assessment seems unfair as more Overall Scores are defined by the Comprehension score than any other profile. Is it too difficult?*	While it is true to say that the Comprehension score is more likely to define the Overall Score than that for any other profile (see ***Report 04 – Summary of TEA Tests conducted & Candidate Performance*** for information about TEA results and trends in profile marking), the TDT maintain that there are several plausible reasons for this trend: • Five of the ICAO Descriptors assess candidates' productive ability, whereas only one (Comprehension) assesses their receptive ability. • A candidate with Level 4 productive skills can control what he produces: this is not true of comprehension since, during a test, the candidate has no control over the structures, vocabulary, accents, etc. they will be exposed to. • The wording of the Comprehension Descriptor at Level 4 is not aligned with the other five Descriptors. For example, the Level 4 description for Level 4 Structure refers to "errors may occur". • Individuals may not have been exposed to the range of accents they experience during the test.
About Part 1…	
TEX: *Occasionally, Part 1 will last for only 2 minutes. This does not seem to be enough time to assess the candidates' ability to talk about familiar, common, concrete and work-related topics.*	The need for standardised interlocution behaviour meant it was difficult to manage this issue in cases where candidates did not *want* to speak in longer turns. The TDT recognised this issue and it was agreed that Part 1 question sets could be extended and that additional, scripted prompts could be added to each set to be used when candidates gave very short answers. As with all test changes, it was recognised that these improvements would require extra instructions and training, where necessary, for interlocutors. Trialling of the new, longer sets concluded that candidates were speaking for more time on common, concrete, work-related topics, eliciting more assessable language than before. Furthermore, the larger question sets allowed for a greater breadth of content, improving the validity of the test. For more information, see ***Report 08 – Item Development & Version Content***.
TU: *The candidates don't always want to give long answers –*	As above, interlocutors cannot force candidates to speak in oral testing. And

they are used to giving clear, concise answers. The interlocutor should be able to prompt them to go into more detail.	it is important for reliability's sake that interlocutors behave in a standardised manner. The extra, scripted prompts helped to overcome this criticism in part. It was agreed that further guidance (to candidates) should be given through additional information in the *Notes for TEA Candidates* (see ***Report 15 - Preparing for TEA***) and the publication of a complete test on the TEA website.
SME: *Questions designed for commercial pilots/ATCOs do not work for private pilots or student pilots.*	Question sets for 4 distinct groups of candidature were piloted and added to Examiner Handbooks: Commercial Pilots/ATCOs, Private Pilots, Ab-initio Pilots, Student Controllers. For more information, see ***Report 08 – Item Development & Version Content***.
About Part 2 generally….	
TU: *Why do our candidates need to listen to Chinese accents during the test when they never hear Chinese accents during their work?*	In designing a test suitable for global purpose for candidates who work in international contexts, and to test comprehension of "a range of speech varieties (dialect and/or accent) or registers', a balance of international accents was required. It is the TDT's position that any international pilot or ATC working in an international context could be exposed to any international accent, whether through routine or non-routine events. For more information about the development of the test recordings see ***Report 08 – Item Development & Version Content***.
ELE/TEX: *The concept of candidates being able to utilise 'clarification strategies' is not entirely accurate in Part 2 – they can only ask for the recording to be replayed but they can't say, for example "What does slippery mean?" and get the answer from the interlocutor.*	There are limits to how authentic a language test can be and, in this case, allowing interlocutors to clarify during Part 2 would spawn practices too varied and unreliable for TEA scores to be considered reliable. It is vital that test interlocution is standardised as far as possible since non-standard behaviour may unfairly benefit or penalise a proportion of candidates. In terms of Part 2, authentic 'clarification strategies' are not viable since they could easily compromise the assessment of Comprehension. In other parts of the test, authentic interaction is permitted.

About Part 2A….	
LTE: *While an ability to read-back or paraphrase short utterances is clearly an aspect of listening comprehension, I cannot say it fully tests comprehension as a 40-item written listening paper might.*	For this testing context, the TDT maintain that: • 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. • From *9835*, under Management of the Dialogue, the following communicative language functions are listed: •Relay an order (C) •Relay a request to act (P) •Relay a request for permission (P) While Part 2A is focussed primarily on assessing candidates' ability to comprehend short messages in non-routine situations, in a productive sense, candidates are demonstrating ability to summarise, read-back or paraphrase and relay that information • Although potentially considered 'purer' tests of comprehension, separate listening tests 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). In respect of this criticism, trials were conducted using recorded, short, question prompts that followed the text and challenged candidates to firstly process the details of the text, then the question, before responding. They proved unsuccessful since the candidates were either over-loaded with information to process, or confused. For more details, see ***Report 03 - Description of Tasks & Instructions***.

TEX: *The interlocutor prompt "What's happening?" does not always lead every candidate to fully explain what they understood.*	To attempt to overcome this problem, the TDT took 2 approaches. The first was to research the effect of different prompts (see ***Report 03 - Description of Tasks & Instructions*** for further details) and adopted new task instructions and prompt. The second, in complementing the first, was to improve the *Notes for TEA Candidates* so that candidates had a better understanding of what was expected of them before taking the test (*see* ***Report 15 - Preparing for TEA***).
SME: *At work, I can make notes. Why can't I make notes while I'm listening to the recordings?*	Although it may improve the face validity of the test, the TDT maintain that making notes should not be permitted for the following reasons: • 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. Note-taking would compromise this. The tasks were not intended to place anything other than minimal and acceptable demand on phonological short-term memory. • Simultaneous note-taking of short texts could impact negatively on performance (see Hale & Courtney, 1994). • Discussions with SMEs revealed that in non-routine, potential emergency situations, there is little or no time to make notes. • Questions over security of test materials would arise as note paper would have to be shredded after each test.
SME: *Part 2A is too short to be able to fully measure comprehension of international accents…*	The TDT conducted trials based on a 20-item Part 2A sets. Both statistical analysis and feedback from examiners and candidates suggested the approach was inappropriate. The former showed that there was no statistical difference in the performance of candidates on 10-item and 20-item sets (see ***Report 03 - Description of Tasks & Instructions*** for further details), and opinions of the examiner and candidate experience were negative. It was decided that adding nothing except test time for the sake of face validity was inappropriate.

SME: *Item 4 is not operationally correct – an ATC wouldn't respond like that in this situation.*	Items triple-checked by operational personnel for authenticity – firstly in the item-writing stage, then during the trialling stage. Disagreements between personnel do occur quite frequently but – where a disagreement occurs during the trialling – a third independent assessment is made about whether the item is potentially obscure / unfair and should be dropped. For more information, see ***Report 08 – Item Development & Version Content***.
TEX: *En-route controllers should only have to listen to recordings related to en-route situations.*	Tests aimed solely at commercial pilots or en-route controllers, would not be appropriate for reasons of language (see above) and practicality - producing a test specifically for en-route controllers would only be of use if the candidates remained in that position for period their test scores were valid. Stakeholders did not want to consider extra testing as a consequence of re-licensing making the concept impractical for the industry.
TEX: *If a candidate doesn't give me a full or clear answer to the situations, why can't I question him or her further to probe comprehension more thoroughly?*	Standardised interlocutor behaviour is a crucial aspect of oral proficiency testing – in terms of both validity and reliability. The TDT consider examiner 'flexibility' in Part 2 of the TEA to be inappropriate since this section is primarily used to assess Comprehension and variability in interlocutor prompting could lead to unfair and inaccurate rating.
SME: *It feels a little bit like a hearing test…some of the recordings are not clear enough and that feels unfair since I never have problems hearing speakers during radio communications.*	Part 2A attempts to replicate authentic non-routine (possibly emergency) situations that include a degree of stress and difficulty. Background noise was initially used to add authenticity and challenge to the recordings. The TDT agreed that any factor that was potentially distracting to a candidate and could be a construct-irrelevant factor in the assessment of listening comprehension should be amended. General feedback from operational personnel suggested that radio communications are typically very clear but that the 'turn' is indicated by the 'click' of the microphone button. Different approaches were trialled involving a variety of 'tinny effects' and radio 'clicks' and clear approval was given by the trial population to one method which was then adopted

	into the newer test materials.
LTE: *The accents seem inauthentic – while actors can produce professionally delivered recordings, the test would be more valid if an authentic range of accents were used.*	The TDT recognised this was an issue in early TEA versions. A selection process led to the employment and training of 25 voice-recording artists to produce future Part 2 materials.
TU: *Student pilots and controllers might not know the technical terms used in the test items.*	The TEA is not designed to be a placement or diagnostic test. If airlines and ANSPs put forward student candidates, it is the TDT's understanding that they are at the end of their training and, therefore, proficient in the necessary terminology for work in an international aviation context.
TEX: *Although the comprehension assessment method is more systematic than before, I'm not sure about the benefit of awarding half-marks. If they understand 50% of the recording, they have not understood enough to say they have comprehended and adding half-points to the whole points means they could reach the next ceiling.*	The TDT team reacted to this by conducting a revision of the assessment of Part 2A of the TEA in which marks (points) were only awarded for wholly understood responses. An analysis of the newer approach to the older resulted in a new system met favourably by raters. For more information, see – ***Report 10 – Establishing Comprehension Score Ceilings for TEA Version 2010***.
About Part 2B….	
LTE: *The two-part format sometimes sounds inauthentic when the first part would not naturally precede the second part. Although this doesn't affect the nature of the task, it could 'throw' a candidate since the ability to anticipate, based on background knowledge of the way things work naturally, plays a part in listening comprehension ability.*	One suggestion was to split the 3 2-part items into 2 separate tasks of 3 independent items – with candidates Asking Questions (new Part 2B) and Giving Advice (Part 2C) independently of each other. Items were written and trials were conducted (see ***Report 08 – Item Development & Version Content***) and the feedback suggested that the new format was less confusing than the previous one. It was hence adopted into TEA Version 2010.
LTE: *In terms of assessment, it is difficult to say that a candidate has comprehended the situation if he only gives generic responses.*	This is true, and to some extent raters have to use some personal interpretation in this part of the test. Open discussions with candidates during trialling led to adjustments that attempt to overcome this potential problem:

	• Clearer pre-test instructions in the *Notes for TEA Candidates* to indicate that there are no correct answers, but that candidates should show they understand the situations by giving relevant responses. • Clearer task instructions to indicate that the situations are in a general aviation context.
SME: *If I hear about a health problem, I'm not a doctor so I don't know what advice to give. It would be better if I had to give advice about technical situations that I am sure about.*	The test is not designed to measure operational knowledge or competence. Trialling showed that operational situations elicit operational language which cannot be assessed by the ICAO Rating Scale. The two adjustments described above were also intended to overcome such candidate doubts. One of the guiding principles in developing test items for Parts 2B & 2C was for item writers to ask themselves whether the candidate would be likely to be able to respond to items easily, sensibly and fully in his own first language. In that way, designing items which elicit suitable language samples while conforming to the focus of a 'broad' work-related context is best managed.
SME: *One or two of the situations are not the type that pilots and controllers would typically have to deal with.*	In trying to elicit plain English in an aviation context, and in order to avoid eliciting procedural responses as much as possible, some of the "unexpected situations" in this part of TEA have to be de-contextualised. That is, although they are clearly relevant to language within an aviation context, it may not be obvious to the candidate that the language they are producing (language of problem-solving – asking questions, offering solutions, seeking clarification) is directly relevant to their operational position.
About Part 3….	
TEX: *Some candidates don't know what to say about the pictures.*	The TDT wanted to give candidates the best opportunity to demonstrate their language proficiency. To this end, clearer information about what was expected in each stage of the test were added the *Notes for TEA Candidates*.
TEX: *Private pilots should not have to describe pictures related*	Since the pictures depict situations of a general aviation nature, and

to commercial aviation.	considering the appropriacy of eliciting plain language in a broad aviation context, the TDT considered this opinion incorrect.
LTE: *If the task was changed to describing and comparing two pictures – as in other established oral language tests – rather than one, there could well be a greater variety of language elicited and the validity of the test content would increase.*	The TDT agreed that this task concept could work well and initial trialling proved as much – the language elicited was more varied in terms of both functions and domains. The biggest problem was the wording of the task instructions but candidate input helped to clarify the best way forward before it was adopted into TEA Version 2010. For more information, see ***Report 08 – Item Development & Version Content***.
TEX: *The content of the discussion topics is too broad – pilots do not know what to say about some of the topics, it's not their job.*	The topics are based on a 'broad' view of work-related context.. One of the guiding principles in developing test items for Part 3 was for item writers to ask themselves whether the candidate would be likely to be able to respond to the question easily, sensibly and fully in his own first language. In that way, designing items which elicit suitable language samples while conforming to the focus of a 'broad' work-related context is best managed. The purpose of communicative tasks in this context is to encourage interaction. The TDT recognised that individuals might not want to give 'incorrect' answers, such is the nature of precision in aviation professional's working life. To this end, clearer information about what was expected in this part of the test was conveyed through the *Notes for TEA Candidates*. The TDT wanted to make it clear that there were not 'correct' answers to the questions and they were simply to prompt interaction. For more information, see ***Report 08 – Item Development & Version Content***.
TU: Security is a big concern for us. What is there to stop a pilot simply 'photo-shopping' a genuine TEA certificate, changing the data and presenting it to us as a genuine certificate?	With this in mind the TDT implemented an online 'lookup' facility for employers, CAA's and other stakeholders to check if the hard-copy certificate presented to them is genuine or not. The certificate number and passport number of the candidate are entered online and the data displayed should match the hard-copy certificate. Candidate privacy is maintained as the certificate numbers always contains random numbers and so cannot be guessed. (See ***Report 14 - TEA Security & Administration*** for more detailed information.)

References

Alderson, J. C. (1993). *Judgements in language testing*. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp.46-57). Alexandria, VA: TESOL.

Alderson, J.C., Clapham, C., & Wall D. (1995). *Language Test Construction and Evaluation*. Cambridge: CUP.

Fulcher, G. & F. Davidson. (2010). *Language Testing and Assessment*. Routledge: New York

Hughes, A. (1989). *Testing for Language Teachers.* Cambridge: CUP.

Hale, G.A., & R. Courtney (1994) *The effects of note-taking on listening comprehension in the Test of English as a Foreign Language.* Language Testing (March, 1994)

The Official Journal of The European Union, November 2011.