MAYFLOWER COLLEGE

Test of English for Aviation

Research notes

MAYFLOWER COLLEGE | RADFORD ROAD PLYMOUTH PL1 3BY DEVON UK
Tel: ++44 - 1752 - 673784 Fax: ++44 - 1752 - 671537 info@aviation-english.com www.aviation-english.com

Executive Summary

This research investigates the performance of the Test of English for Aviation (TEA). As there is no agreed single instrument to measure the reliability and validity of a test, a number of studies have been conducted which, taken together, provide a picture of TEA's performance.

The Test of English for Aviation (TEA) assesses the plain language proficiency of air traffic controllers and pilots in an aviation context. The test lasts approximately 20 minutes and elicits language that can be assessed according to the ICAO Language Proficiency Rating Scale. TEA was designed and constructed according to the recommendations of ICAO Document 9835. The test designers drew on their many years of experience of teaching Aviation English, administering IELTS examination centres for over 10 years, examining for IELTS, examiner training for IELTS and the pre-testing of IELTS papers on behalf of Cambridge ESOL.

Following the creation of test specifications a number of tasks were constructed and trialled with operations personnel. After these tasks were analysed the final version of TEA was produced. A study was conducted to establish how easily candidates were able to engage with the tasks, and whether the language required by the test reflected work-related communications. An introspective data-gathering exercise was carried out (see Cohen 1984, Faerch and Kasper 1987 and Grotjahn 1986). The results were positive with most candidates reporting that the language required by TEA was authentic and that they were able to engage in each of the tasks.

Observation Checklists were used to investigate whether the language functions the test was intended to elicit were actually produced by candidates. Observation Checklists were developed by Cambridge ESOL to validate the First Certificate exams, and are so-called because examiners watched video recordings of the tests. They have also been used by IELTS with tape recordings of tests. For a further explanation see Saville and O'Sullivan (2000), Brooks (2003). Observation Checklists for the TEA test were developed and trialled over a number of months. The checklists were applied to 34 tests and the results were analysed to measure the language elicited from candidates between levels 3-5.

The results show that the test is performing largely as it was intended to. The majority of candidates produce the functions expected in each task, although there are some differences between levels. For example, level 2 candidates tend to express misunderstanding more frequently than a level 5 candidate. The checklists are now used to monitor the performance of new items in trial versions.

TEA Examiners undergo rigorous **training** and ongoing monitoring. The performance of senior examiners is monitored regularly partly through group standardisation sessions and also by conducting reliability studies. Inter-rater reliability is checked by selecting tests which had been rated by the group of senior examiners and asking examiners to re-rate them individually three months later. The results are presented using the Pearson Correlation Coefficient and the Average Absolute Difference in the marks awarded. The overall mark awarded by an individual examiner is compared to the mark previously awarded by the standardisation group. If the examiner awards the same mark as the standardisation group, the Absolute Difference is 0. On the other hand, if the standardisation group gave a level 2 and the examiner gave a level 4, the Absolute Difference would be 2. Therefore, the closer the Difference is to 0, the more reliable is the examiner's performance. With the Pearson Correlation Coefficient, 1 indicates perfect correlation; therefore the closer the results are to 1, the more reliable an examiner's marks.

Inter-rater reliability appears to be very high. The Absolute Difference in marks awarded by all the senior examiners in their individual standardisation sessions has been collated and the Average Absolute Difference is 0.2. When expressed using the Pearson Correlation Coefficient the result is 0.82.

Intra-rater reliability is also investigated to determine whether individual examiners are consistent over a period of time. Examiners are asked to re-rate tests which they had assessed at least three months previously. The results of this phase of research are encouraging. The Average Absolute Difference between the first and second ratings is 0.13 and the Pearson Correlation Coefficient is 0.94.

Concurrent validity was investigated by comparing the results awarded to candidates in the TEA test with assessments of their levels made by teachers who had worked with them. The study was carried out with 50

students from Romania, Kyrgyzstan, Italy, Spain and Brazil. Each group of students attended a 4-week (100 hours) English for Aviation course to help them achieve ICAO level 4. The teachers working with them were familiar with the ICAO Language Proficiency Requirements and the application of the ICAO Rating Scale, although they were not TEA examiners. The results demonstrate a very high degree of concurrence between the teachers' assessments and the TEA results. When expressed using the Pearson Correlation Coefficient the degree of concurrence is 0.82 and the Average Absolute Difference between the predictions and the marks awarded is 0.26.

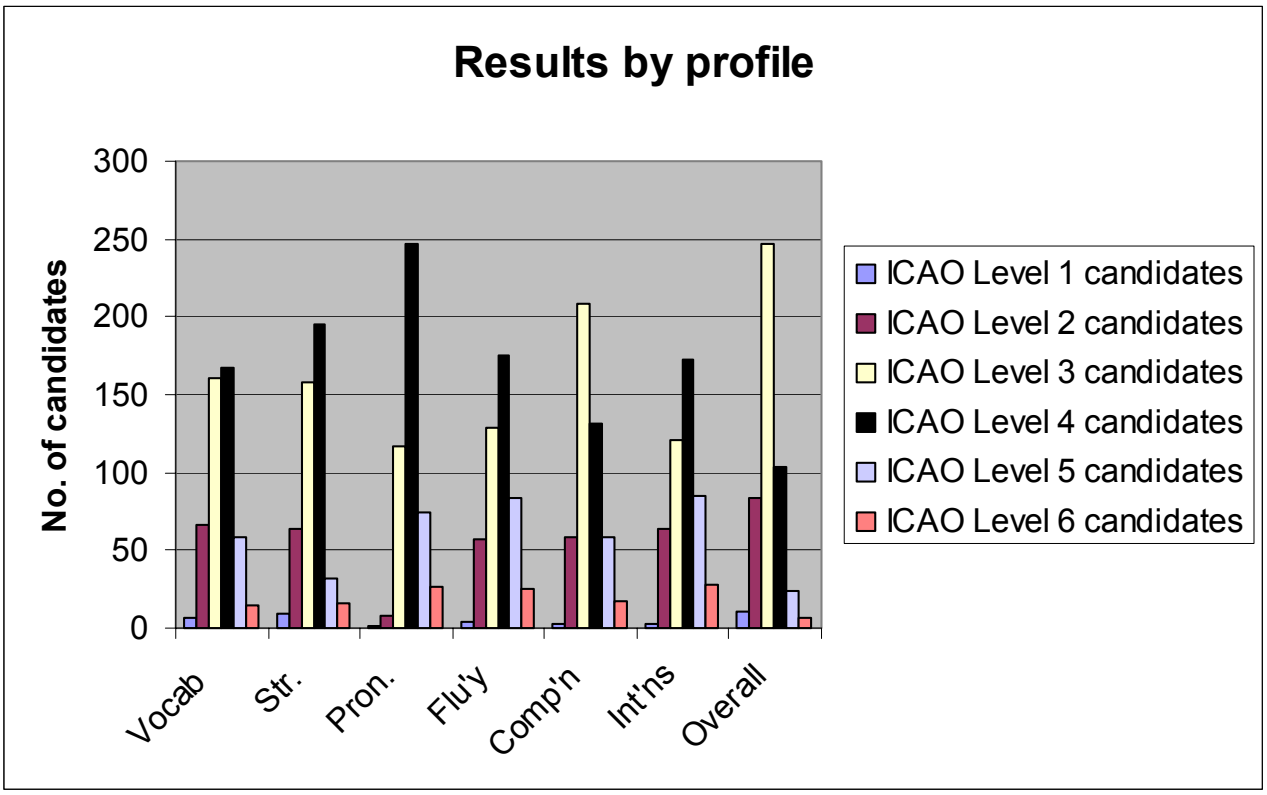
This study was augmented by a similar exercise carried out with a number of aviation training institutes who had used TEA. Teachers at these institutes were asked if they were able to say if the marks their students were awarded by TEA matched their predictions, based on their observations of those students in the classroom. Whilst we do not have access to specific data for each of the candidates who took the test, partly due to the sheer numbers involved, the feedback from teachers was that, generally, the marks awarded reflected their expectations. Stakeholders also reported that they felt the test presented authentic situations and elicited appropriate responses from candidates.

Word count: an analysis of the transcripts reveals that a TEA examiner utters an average of 486 words in every test (this count does not include the items / questions which are CD-Rom based). As an examiner's prompts are heavily scripted the standard deviation on this average is relatively small (even allowing for candidates' opportunity to request clarification, paraphrasing, etc). The average word count for speech produced by candidates is 1057 for Level 2 candidates, rising to 1849 for Level 5 candidates.

The validity of a test and the reliability of its results are easily compromised by **security** breaches. Some of the threats include impostors taking the test and materials being removed from test centres. TEA has a number of systems in place to reduce these risks, including stringent candidate identity checks, photographing candidates before they enter the test room, and the centralised production of certificates. To date, there have been no recorded instances of security breaches.

TEA will continue to be evaluated using the methods outlined in this report. We would also like to explore other areas, in particular the creation of a lexical corpus and parts-of-speech analysis. We would also like to explore the possibility of greater involvement from the wider academic community.

This table shows the overall ICAO level of 474 TEA Candidates



Test Construct

Overview

The Test of English for Aviation assesses the plain language proficiency of air traffic controllers and pilots in an aviation context. The test lasts approximately 20 minutes and elicits language that can be assessed according to the ICAO Rating Scale. TEA was designed and constructed according to the recommendations of ICAO Document 9835.

Test Development Process

The test design process began in September 2004. Tasks and items were developed in accordance with ICAO Document 9835, the *'Dominant Communicative Functions In Radiotelephony Communications'* and the ICAO *Priority Lexical Domains* (Appendix B of Document 9835). Since 1992 Mayflower College has been working with operations personnel in a teaching capacity. Item writers and test designers have extensive experience of English for Aviation and we were able to trial materials at each stage of development. The item writers also had many years experience in trialling and pre-testing materials for the IELTS exam.

The test designers produced the following general specifications.
The test should:

- Be suitable for all types of pilots and ATCOs
- Elicit language to assess ICAO levels 1-6
- Test plain English, not Standard Phraseology
- Allow production of multiple and standardised versions
- Be sufficiently secure for this high-stakes environment
- Be relatively economic to administer

The reasons for these objectives were as follows:

"Suitable for all types of pilots and ATCOs"

Appendix B of Document 9835 lists the communicative language functions associated with aviation. The vast majority relate to both pilots and air traffic controllers, rather than just one of these professions. Given that the underlying aim of the LPRs is that both pilots and ATCOs shall communicate effectively in both routine and non-routine situations it seemed logical to produce a test suitable for all personnel. It is true that different types of controllers perform different roles at work, and therefore occasionally have different language needs. However, it was felt that aviation safety could best be promoted by covering the general needs of the professions. Producing a test specifically for en-route controllers, for instance, would only be of use if the candidates remained in that position for the period that their test scores were valid. If, in the three years after the language test, they moved from en-route to ground control (as is often the case) their language score might not be appropriate.

"Elicit language to assess ICAO levels 1-6"

Stakeholders from the industry had asked for a test which could assist in training needs analyses and the test designers wanted to encourage positive washback as far as possible. (Unfortunately this is notoriously difficult to measure - see O'Sullivan 2005). If stakeholders were to design and implement effective training programs they would need an accurate assessment of the level of their personnel. It would have been unhelpful to design a test that only placed candidates at levels 3 and 4.

"Test plain English and not Standard Phraseology"

ICAO Document 9835 clearly points to plain English in an aviation context. The document makes it clear that whilst standard phraseology is fundamental to aviation safety, there are many non-routine situations that require other communication strategies. The ICAO Rating Scale was clearly designed to assess plain English, rather than phraseology (for example by measuring paraphrase, idioms, register). ICAO elaborated on this point in June 2006:

"Just as testing of ICAO phraseology cannot be used to assess plain language proficiency, neither can English language proficiency tests be used to test ICAO standardized phraseology.
*It is acceptable that a test contain a scripted test task in which phraseology is included in a prompt. The test task may be used as a warm up or an ice-breaker and elicit a **plain language** response from the test taker."*

It is our view that the testing of Standard Phraseology needs to be assessed by operational experts using a different set of criteria (not the ICAO Language Proficiency Scale).

The test designers also wanted to ensure, as far as possible, that the test measured only language proficiency, rather than intelligence, logical thinking, or operational knowledge.

"Allow the production of multiple and standardised versions"

The scale of the 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, and need to be replaced regularly to maintain the confidentiality of materials (and by extension, the reliability of results). The tasks needed to be standardised to ensure that all sets were of a similar level of difficulty and elicited similar language.

"Security in a high-stakes environment"

This aspect of the test design related more to the systems of administration and delivery than task design but was still a factor that the test designers had to consider. With aviation safety and peoples' livelihoods at stake, there was always the risk of candidate collusion or cheating. This would be reduced by having multiple versions of the test (see above), and 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. Initial explorations into the possibility of computer-based tests were quickly discarded because these would not be secure enough. (The security of TEA is analysed below.)

"Economic test administration"

The overriding concern of the test designers was to help promote aviation safety through the construction of a valid, reliable testing system. The key notion however is "Affordable Safety" – there is always a trade-off between Safety and Cost. The test itself was only part of the solution to the problems of poor communication within the industry – indeed, in many ways, was only the start of the solution. Stakeholders would need to invest heavily in appropriate training programs in order to help their personnel achieve ICAO level 4. It was critical then that the test offered an affordable means of assessing personnel, in order that airlines and ANSPs could maximize their training resources.

Initial Trial Versions

A number of tasks were developed and trialled with operations personnel. These tasks included:

- Work-related interview
- Description of the events presented in a video
- Problem-solving role-play
- Discussion of aviation topics

These were trialled with personnel from France, Lithuania and the UK and analysed to determine:

- Whether the language elicited could be assessed according to the rating scale
- How easily candidates could engage in the tasks
- Whether multiple, standardised versions could be produced

Candidates who helped trial these tasks were interviewed after the test by a teacher, and asked how they felt about the task, how easily they felt they could participate in it, and what they thought the purpose of the task

was. As this research was qualitative rather than quantitative it is not possible to collate the results for this report.

Work-related interview

Candidates were asked a series of questions on common, concrete, work-related topics. All candidates were able to respond to these prompts to some degree, enabling a meaningful sample of language to be elicited. Candidates reported that they felt comfortable with this task, and it was retained in the final prototype of the test.

Video

Candidates were shown a 30-40 second video of an emergency situation. They were asked to describe the video as it was played, and then to summarise what they had seen afterwards. This task was designed to test candidates' ability to respond to non-verbal cues, convey information and paraphrase, as well as use an appropriate range of structures. Candidates were not given any preparation time, in order to simulate a stressful work situation. Trials of this task showed that it discriminated against lower-level candidates who lacked the vocabulary and fluency to describe the situation quickly. They tended to lapse into silence, indicating a Level One user, whereas in fact their language proficiency was higher than this. Many candidates reported that they found the task difficult to complete due to the time pressure.

Problem-solving role-play

In this task the examiner played the part of somebody with a problem, which the candidate had to resolve by finding information and giving advice. The examiner presented the candidate with limited information about the problem in order to encourage the candidate to ask questions. In addition, the examiner would convey irrelevant or unhelpful information. The candidate was required to use their interactive skills to control the conversation and select relevant information from the examiner. In the trial versions it was found that that the amount and type of language elicited from candidates varied considerably and it seemed to be more a test of cognitive strength than linguistic proficiency. In particular, candidates did not always feel able to interrupt when the examiner was relaying useless information (distracters). It was felt that this was partly a cultural factor which might discriminate against certain nationalities. This task also required imagination and creativity from the examiner, which made it hard to produce a standard rubric.

Discussion

In this task candidates were asked to discuss more general aviation topics with the examiner. Initial trials were conducted with examiners adhering to a standardised rubric, as in other trial tasks. It was found that it was difficult to grade this appropriately for all levels of candidate. Candidates with a lower level of English often reported that they were unable to respond to the task because they did not fully understand the questions and examiners felt that the standardised rubric was too restrictive at higher levels. Consequently, a second version of this task was trialled giving the examiner greater flexibility. Rather than using a standardised script, a series of prompts were developed which the examiner could use to develop the discussion with the candidate. This was trialled and found to be a great improvement.

Second stage trial tasks

New tasks were constructed to attempt to simulate situational complications, encourage problem solving and expose candidates to a range of international accents. Again, these tasks were trialled with operations personnel and examiners who were then asked to give their views on the tasks.

Describing a route

A new task was developed in which candidates were given a map and asked to describe the route between two places. This task was designed to test extended fluency and the candidates' response to non-verbal cues. Vocabulary was tested because candidates had to include as much information about the travel conditions as possible. The task was also designed to reflect the Priority Lexical Domain of Geography and the Critical

Components of Conveying Information, Giving Advice and Issuing Warnings. Candidates were given a chance to study the map before speaking and had limited time in which to complete the task.

The initial trial proved successful with candidates reporting that they felt comfortable attempting this task as it was work-related, and examiners reporting that they were able to elicit appropriate language from the candidates. One concern that arose was that candidates did not always understand the symbols on the map and item-writers considered using written labels. However, this would have required a degree of reading comprehension, which would discriminate against some candidates, and was not in keeping with the requirements of 9835 that reading should not be tested. Instead, examiners were instructed to disregard misinterpretations of symbols, provided candidates were able to convey meaning.

It was also easy to produce multiple versions of this task and it was retained in the prototype of the test.

Interactive comprehension

Candidates were played a series of statements delivered in a range of international accents. Candidates were asked to respond by asking questions to get more information about what was happening, and then to give advice to help resolve the situation. The statements were drawn from pilot-controller communications in non-routine situations, documented in CVR transcripts and aviation publications. This task was designed to test comprehension of a range of international accents, the use of clarification strategies and the immediacy and appropriacy of response.

Initial trials of this task showed varied results. Many candidates were unclear about what they were supposed to do and their performance consequently suffered. The task was piloted again with an unrated example to demonstrate the task, which eliminated this problem. A number of candidates also reported that they found it difficult to respond to all of the recorded statements, as they did not understand all the accents. As this was one of the fundamental skills being tested in this task the test designers felt that it was appropriate to retain it in the final version of the test.

Final version prototype

Following the initial trials of different task types, the following test format was selected.

Part 1- Interview
Part 2 –Description and directions
Part 3 – Interactive Comprehension
Part 4 –Discussion

This test format was then trialled with 20 operations personnel from Romania, Italy and France. These candidates were selected for two reasons: (i) they wanted to have an idea of their ICAO language level and would therefore take the test seriously, (ii) as they were studying at Mayflower College, examiners would be able to compare their performance in the test with the language they produced in class to see if the test gave a fair measurement of their performance.

Candidate Moderation Survey

The test design team particularly wanted to investigate how easily candidates were able to engage with the tasks, and whether they felt the language required was authentic. An introspective data-gathering exercise was carried out (see Cohen 1984, Faerch and Kasper 1987 and Grotjahn 1986). Questionnaires were developed to investigate how candidates felt about the test. These questionnaires were trialled with a small group of aviation students to ensure that they were non-directive and that the language used was accessible to all levels. The results of the questionnaires were positive, with the majority of candidates reporting that they felt comfortable during the test and that the tasks reflected their professional lives. A small proportion of candidates disliked the test because it did not include standard phraseology. However, in this respect the test construct met the requirements of Document 9835. Some minor changes to the rubric of the test were made after the questionnaires were analysed, but otherwise the basic format remained unchanged.

Whilst this sort of investigation is qualitative, some results were quantified:

	Long	Short	OK	Don't know
What comments could you make about the length of the test?	0	4	16	
	Y	**N**	**N/A**	
Were the instructions during the test clear?	19	1		
	Y	**N**	**N/A**	
Were there any times when you didn't know what to do?	3	17		
Part One				
	Easy	**Difficult**	**Mixture**	**N/A**
Did you find the questions easy or difficult?	17	3		
Part Two				
	Easy	**Difficult**	**Mixture**	**N/A**
Did you find the questions easy or difficult?	9	1	8	2
Part Three				
	Easy	**Difficult**	**Mixture**	**N/A**
Did you find the questions easy or difficult?	12	1	5	2
Part Four				
	Easy	**Difficult**	**Mixture**	**N/A**
Did you find the questions easy or difficult?	9	4	7	
General	**Y**	**N**	**N/A**	
Do you think it is a good test?	14	3	3	

This study was repeated over a period of three months with the finalized version of the test. Candidates included ATCOs, pilots and instructors from Romania, Russia and France. The results were very similar to those generated by the initial study, with most candidates saying that they felt it was a good test for the ICAO requirements, and a minority reporting that they felt standard phraseology should also have been included. Crucially, candidates reported that they felt they understood what was required of them throughout the test and that, generally, they felt the tasks set were achievable.

	Long	Short	OK	Don't know
What comments could you make about the length of the test?	1	5	36	4
	Y	**N**	**N/A**	
Were the instructions during the test clear?	46			
	Y	**N**	**N/A**	
Were there any times when you didn't know what to do?	5	40	1	
Part One				
	Easy	**Difficult**	**Mixture**	**N/A**
Did you find the questions easy or difficult?	40	2	3	1
Part Two				
	Easy	**Difficult**	**Mixture**	**N/A**
Did you find the questions easy or difficult?	28	11	5	2
Part Three				
	Easy	**Difficult**	**Mixture**	**N/A**
Did you find the questions easy or difficult?	24	6	8	8
Part Four	**Easy**	**Difficult**	**Mixture**	**N/A**
Did you find the questions easy or difficult?	29	6	3	8
General	**Y**	**N**	**N/A**	
Do you think it is a good test?	36	2	6	2

Revising the TEA Test

(i) Comprehension and Interactions

Results of ongoing research found that while the test was working well, improvements could be made to Part Three. Examiners felt that the existing comprehension part of the test was not performing well enough as it measured both comprehension and interactions simultaneously. This was a direct reflection of the ICAO rating scale, in which the immediacy and appropriacy of a candidate's response is inextricably linked to their ability to comprehend a message. However, the existing test did not fully test comprehension skills alone. Candidates were in effect only required to understand the gist of a prompt rather than specific detail. Their comprehension was measured in terms of their ability to question and advise and not in any more direct fashion.

A new section was piloted in which candidates were played verbal prompts delivered in a range of international accents and asked to explain what was happening in each one. The prompts were derived from authentic pilot-

controller communications and covered specific lexical domains: human, environmental, health and technical. The messages followed the same format in order to standardise the task and allow the production of multiple versions.

The trials of this task showed that it worked well at all levels of the candidate population. Higher-level candidates were able to paraphrase the prompts successfully and lower-level candidates were able to give effective readback. This task was occasionally problematic for candidates who had limited experience of a range of accents but as this is a prerequisite for achieving level 4 this was deemed acceptable. This section was introduced into the new version of the test.

(ii) Description

One of the aims of Part Two, where candidates were given a map, was to elicit a description of the conditions en route, partly to help assess the range and accuracy of lexis and also because this fitted with the priority lexical domains. However, it was noted that a number of candidates were simply giving directions without reference to conditions. The brevity of this type of response often belied a much higher level of language, as evidenced in other parts of the test. Candidate moderation forms indicated that candidates felt comfortable with this part of the test because it was strongly work-related. Although this had initially been viewed as a positive factor, one of the requirements of the test was that candidates were presented with situational complications.

(iii) Picture description

A new task was trialled in which candidates had to describe aviation-related pictures. They were given a short preparation time and were asked to talk for a maximum of thirty seconds. This task was particularly designed to examine vocabulary, structure and fluency, and to see if candidates could respond to non-verbal cues. The trial showed that all candidates were able to respond, although the length of the response and the sophistication of language used naturally varied according to linguistic proficiency. Candidates also reported that they found many of the pictures interesting to talk about, which the test designers felt was another positive point. It was also relatively straightforward to produce multiple versions of this task and it was retained in the new version.

Final Version

The final version of the test is as follows:

- Part One: Work-related interview
- Part Two: Interactive Comprehension
- Part Three: Description and Discussion

Observation Checklists

Overview

Observation Checklists are designed to measure how often candidates use different language functions during a test. They help to show whether a test is eliciting the language we expect it to. Observation Checklists were used by Cambridge ESOL to validate the First Certificate exams, and are so-called because examiners watched video recordings of the tests. They have also been used by IELTS with tape recordings of tests. For a further explanation see Saville and O'Sullivan (2000), Brooks (2003).

The Construction of Observation Checklists

The TEA Observation Checklist was constructed to reflect the functions that the test designers hoped to elicit from candidates. These functions had especially been based on Appendix B of Document 9835 (Communicative Language Functions, Events, Domains and Tasks associated with Aviation). This lists 116 functions of pilot-controller communication. For the purposes of developing an effective and practical checklist which could be applied easily in real time these functions were simplified or collated, as in the examples shown below:

9835
- Describe a state
- Describe a changed state
- Describe an unchanged state

was simplified to "Describe a state" in the Observation Checklist.

9835
- Give an order (C)
- Give an amended order (C) Give a negative order (C) Give alternative orders (C) Give a sequence of orders (C) Cancel an order (C)
- Announce compliance with an order (P) Announce non-compliance with an order (P)

was simplified to

"Orders" (Give an order, state (non-) compliance with an order) in the Observation Checklist.

In this way, 45 of the functions in Appendix B were included in the Observation Checklist, with the majority pertaining to both pilots and controllers, reflecting the design specifications of the test. Other functions, not listed in Appendix B, but which the test designers hoped to elicit were also included. In addition to this, the checklist incorporated further language functions which were neither in Document 9835 nor in the test specifications, but which might nevertheless appear in the language elicited.

First Draft

The first draft of the checklist was trialled with a group of 4 TEA examiners using the same test. The results from the examiners varied considerably, because the checklist was too unwieldy to use effectively.

From this initial trial, two decisions were taken: (1) to simplify the checklist by reducing the number of functions it incorporated; (2) to operate the checklists not just in real-time but also with transcripts.

The checklist was reduced and all functions not intended to appear in the test were removed. Whilst this opened the procedure to the charge that it was not fully measuring the language used by candidates, we hoped to resolve this problem through the use of transcripts. Examiners were asked to map the functions onto the transcripts in order to give a visual impression of how much of the language elicited could be assigned a function. If significant amounts of dialogue were left unmarked, the checklists would have to be revised.

The use of transcripts was a departure from the intended application of checklists in real time. However, literature on the subject states "real-time checklists" were largely intended to reduce the cost of transcribing

tests and studies have found a high degree of concurrence between checklists applied in real-time and on transcripts. (O'Sullivan, Weir and Saville 2002) We felt that the results were therefore likely to be just as valid and that using transcripts (which had already been produced) would, in fact, be easier and less time-consuming for examiners.

Second Draft

The second draft of the checklist was trialled again with 4 examiners and found to be a great improvement in its ease of use. Some disagreement remained over the functions that best described certain features of speech. For example, a candidate describing a typical day was both "providing information on past / present experiences" and "describing a sequence of events". Whilst this suggested it would be hard to achieve perfect calibration, the purpose of the exercise was to investigate whether the required functions were elicited in the test at all. Therefore, where functions overlapped, as in this case, it was not seen as a significant hindrance.

In order to try and achieve greater standardisation, further training materials were produced, with functions mapped onto a test transcript and presented to the group of examiners. This was intended to further categorise the language functions elicited. A further function, "staging", was added, to describe candidates setting the context of the next utterance, or verbalising their thought process.

Checklist Standardisation

A standardisation exercise was conducted after the training session using transcripts of 5 tests. Examiners then individually applied the checklist to one test. The examiners agreed on which functions appeared in the test (16 functions were identified). Of these 16 functions, examiners agreed on the frequency with which 11 of the functions appeared; and the majority of the examiners agreed on the frequency of use of a further 2 functions. There was disagreement on the frequency of two functions: "Check, confirm and clarify"; and "expressing misunderstanding", due to a misunderstanding of the checklist. Some assumed that if a candidate needed to check information then by definition they had not fully understood the message. For the same reason, there was a slightly higher level of disagreement on the frequency with which functions appeared in each task.

TEA Observation Checklist (final version)

FUNCTION	ELABORATION
Providing information	Giving information on past, present experiences
Request information	
Expressing opinion	State what you think about a topic
Elaborating	Elaborate on or modify an assertion
Explaining	Express reasons for an assertion
Preferences	State or ask about preferences
Needs / wishes	State / ask about needs / wishes
Speculating	Suggest an idea; hypothesise; formulate a theory
Describe	Describe a sequence of events; a process, a procedure, the source of a problem, a state, a visual impression; quote rules
Comparing	
Predict	Predict a future action / event; state possible consequences of an action / event
Necessity	State / ask about necessity
Orders	Give an order, state (non-) compliance with an order
Approval and permission	Give or request approval to act
Requests	Request action by another; agree / refuse to act
Offers	Offer to act; refuse / accept offer to act
Advice	Give advice; suggest a course of action or solution to a problem
Checking, Confirming, Clarifying understanding	
Express misunderstanding	
Request repetition	
Summarising	Reiterating or paraphrasing what another has said or giving, readback
Express concern	
Reassure	
Encourage	
Staging	Setting the context of the next utterance, verbalising a thought process

The Method of Applying the Checklist

Only the parts of the test which are assessed according to the ICAO Rating Scale are analysed in the Checklist. Therefore, the introduction (where the candidate confirms their name and presents their identification) is ignored as are responses to unrated example questions.

The checklists only measure the functions produced, whereas the descriptors measure communicative proficiency. So, a poorly constructed sentence ("I am working here for 29 years") might be penalised under the descriptors, but would still be still be recorded as "Providing information" in the Observation Checklist.

Similarly, the checklists do not measure the appropriacy of a response. For instance,

Examiner: "*Tell me about your job. Can you tell me what you do*?"

Candidate: "*I am sitting down*."

The descriptors might penalise this under comprehension, but on the checklist we would record Providing Information.

Stretches of language may contain only one, or several, functions. Compare:

"*I don't understand*." (1 function: expressing misunderstanding)

"*I don't understand maybe you can play it again?*" (2 functions: expressing misunderstanding and requesting repetition.)

This can also occur in longer stretches of language, for example when the candidates are asked to describe an aviation situation. Compare:

"In this photo I see a plane making an emergency situation, he is to make an emergency landing, uh, all the passengers have to evacuate."

The only function is **Describe**

"*This aircraft land in airport with, uh, problem engine, engine is fire, and maybe problem with fuel, and I hope this aircraft is ok with passenger. Ha ha, is it enough?*"
Here there are 4 functions:

"This aircraft land in airport with, uh, problem engine, engine is fire" **Describe**
"*and maybe problem with fuel*" **Speculation**
"and I hope this aircraft is ok with passenger" **Express Concern**
"Ha ha, is it enough?" **Checking**

For ease of use, we have simplified the functions recorded where candidates are describing photographs. Almost all descriptions are likely to include providing information, elaborating and explaining – as this is the very nature of describing something. Therefore, when candidates use these functions to provide descriptions, the only function recorded on the checklist is "Describe". However, where candidates provide other functions, as in the example above, these are recorded.

Examples of the functions used

These examples come largely from candidates who achieved an overall level 2 or 3. Experience shows that it is more difficult to apply the checklists at lower levels because candidates often have poor control of structure and lexis, making the function harder to interpret. At levels 4 and 5, candidates are generally able to communicate with greater precision.

Providing information - Giving information on past, present experiences

"*I'm air traffic controller from ……"*
"I work with a co-pilot and an engineer."

Asking for information

"What's wrong with that woman?"
"Where from the smoke?"

Expressing opinion

"It's very interesting work. I enjoy it every time when I do my work, do my job."
Elaborating

"It's very interesting work. I enjoy it every time when I do my work, do my job."

Explaining

E: And do you get on with your colleagues?
C: Uh yes.
E: Why?
C: We have good relationshipment. (sic)

Speculating

"I see here a plane and a I don't know what these people do on this plane. What more. I think he, they are clean, cleaning this aircraft this plane."

State preferences

"I'd prefer to work in a smaller team because…"

Describe

(i) A sequence of events

"Oh…come to my work maybe at 6.30 am we check all, check all parts, radio, radar uh and etc then I watch a radar and try to control and so on"

(ii) A process or procedure

E: *Do you have to do anything different if there are strong winds for example?*
C: *If we change weather change…it impossible to change…airport…some pilot…waiting…also to airport hangar cleaning our.*

(iii) A visual impression

In this picture I see aircraft has been with fire engine uh I think it make priority landing. And uh and he had collision with runway and that why engine getting fire but I don't see here um what more.

Advice

E: *Give advice.*
C: *Ah if you have some medicine for this nervous, uh nervous this woman and give it to her.*

Checking, confirming, clarifying *(any request for reiteration but NOT a direct request for repetition.)*

E: *Can you describe a typical working day?*
C: *Excuse me? At work or at home?*

Express misunderstanding

"I don't understand this question"

Request repetition

"Say again please."

Summarising

Virtually all attempts at answering part 2a – whether the information is accurate or not.

Express concern

"This aircraft land in airport with, uh, problem engine, engine is fire, and maybe problem with fuel, and **I hope this aircraft is ok with passenger**. Ha ha, is it enough?"

Staging (Verbalising the thought process or setting the context of the next utterance)

E: *Can you describe a typical working day?*
C: *Typical. Typical. Oh, typical - normal! Ok, I report to work at….*

Applying the Checklists

Examiners then applied the checklists to 34 tests. The test candidates were from 5 different countries and their overall levels ranged from 2-5.

Results

The results were analysed to investigate the way that different levels of candidates respond to each task. There are certain functions which one would expect to elicit from a low-level candidate (e.g. expressing misunderstanding or checking, confirming, clarifying) but would not expect to elicit from a higher level candidate. In order to examine the performance of the test, the results of the checklists have been sorted by the mean average of the candidates' level (i.e. taking the combined marks for pronunciation, vocabulary, structure, interactions, comprehension and fluency and dividing by 6). We have not used the overall ICAO level (the lowest of the 6 marks) as this does not always reflect a candidate's general language ability. Many candidates, for instance, are fluent, accurate and intelligible, but have difficulty comprehending a range of international accents.

The process of mapping the functions onto the transcripts showed that the observation checklist was sufficiently comprehensive to assign a function to most of the language elicited from candidates. Exceptions occurred with extremely low-level candidates who produced language so devoid of form that it was impossible to deduce their communicative intent.

The results demonstrate that all levels of candidates attempted to respond to the tasks with appropriate functions and from this it was judged that the test was working well. As expected, certain functions were more prevalent at particular levels. For instance, level 4 candidates were more likely to request repetition of a message than level 5 candidates.

The checklists suggest that candidates are able to respond with the appropriate informational and interactional functions in each part of the test. For instance, in part one, candidates of all levels conveyed information, in part two there were many instances of summarising, checking, confirming, clarifying and giving advice, and in part 3 candidates were able to describe, explain and elaborate.

The following charts show how many times, on average, candidates at levels 2 – 5 used each function during the test.

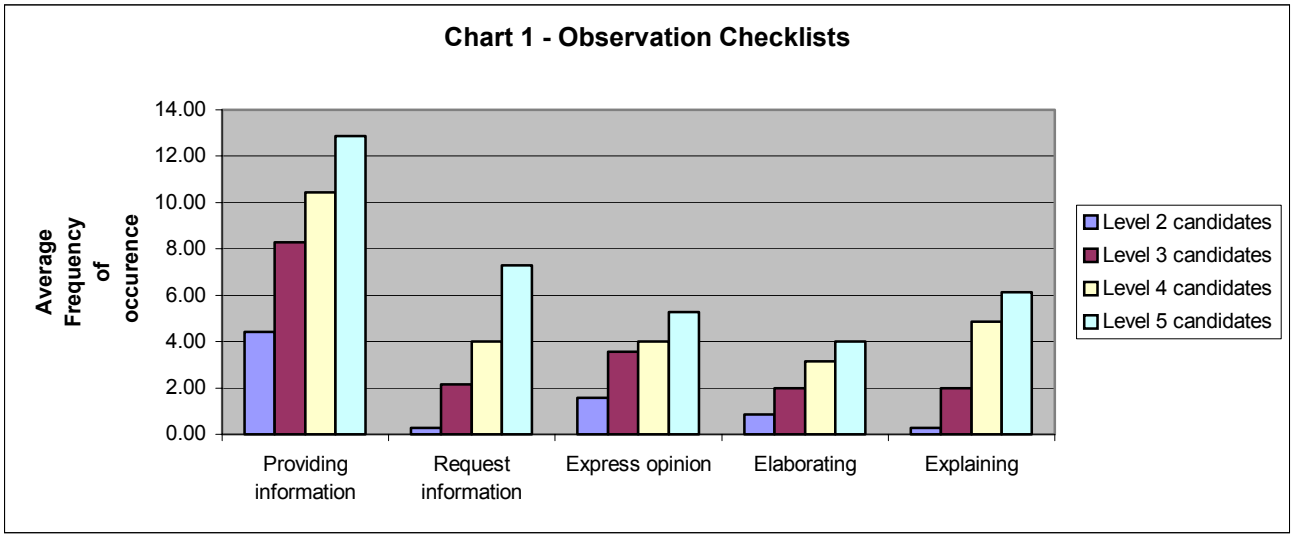


Chart 1: the average frequency of occurrence increases in direct proportion to the overall level of the candidate. Level 5 candidates, who are more fluent and have a greater lexical range are able to provide more information, for instance, than candidates of lower levels.

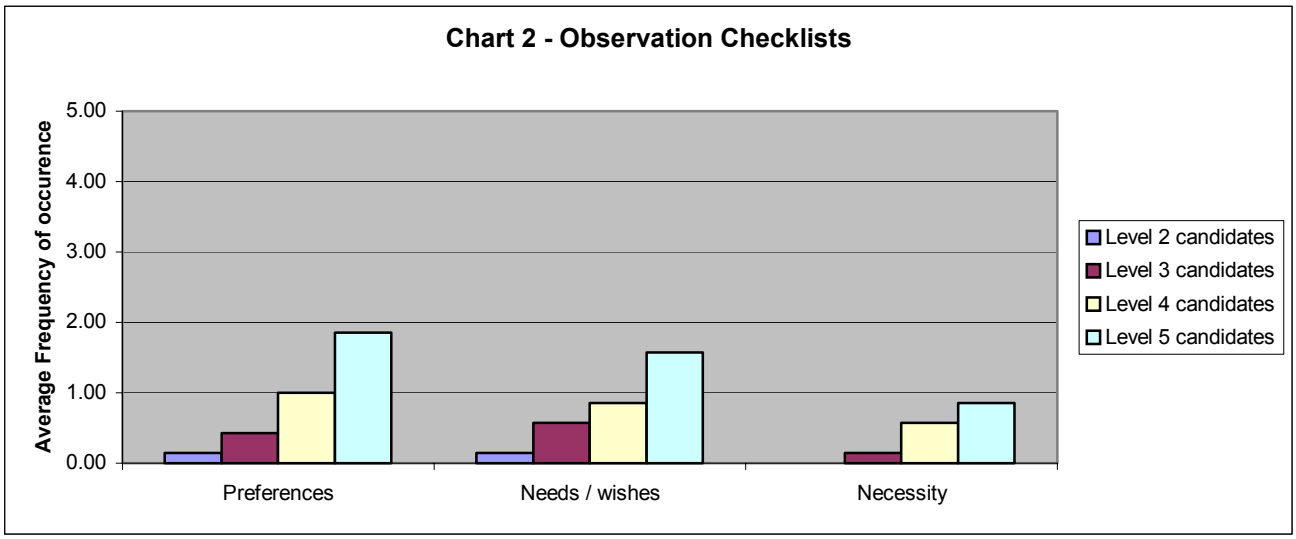


Chart 2: the functions occur less frequently largely because they overlap with other functions on the checklist. For example, a candidate explaining **Preference**s ("*I like night-time flights more than flying during the day*") is **Providing Information**. **Necessity** also overlaps with **Describing a process, and quote rules** ("*I need to go for a medical check ever six months to keep my licence*").

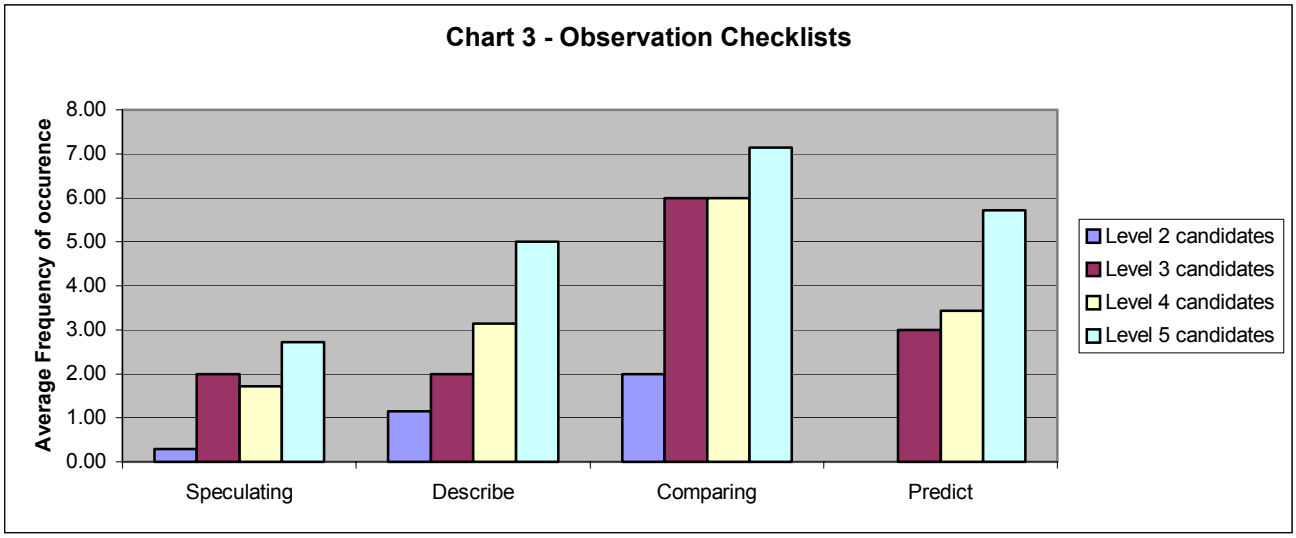
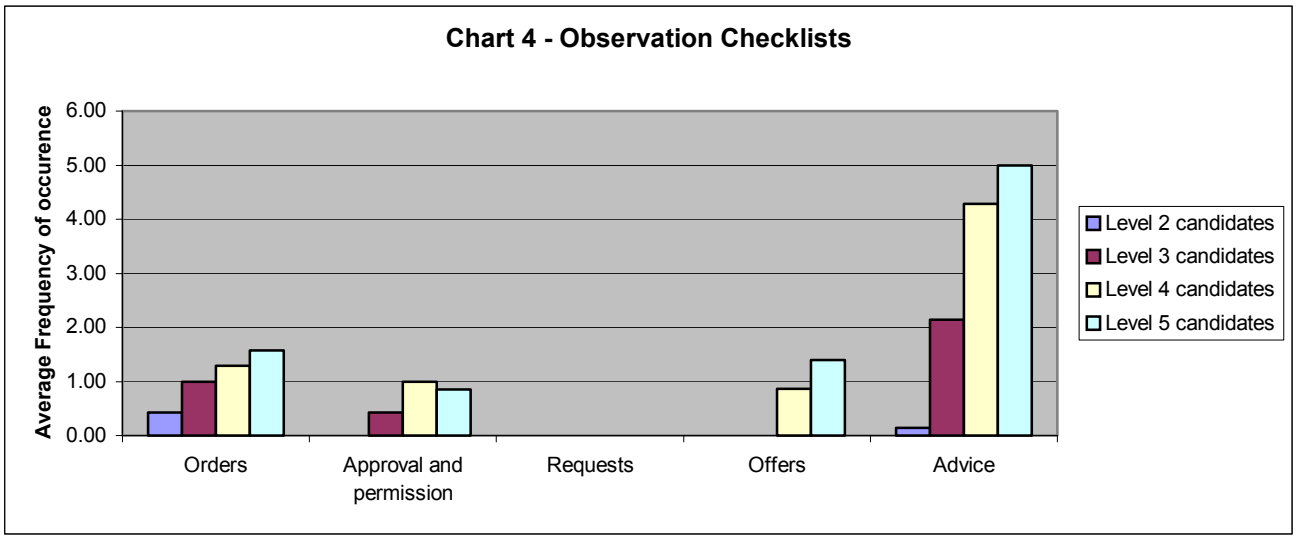


Chart 3: Again the frequency of occurrence shown here is generally in direct proportion to the candidates' levels. The exception is that level 3 candidates tend to speculate more often than level 4 candidates. An examination of the test transcripts showed that this was largely due to responses to Part Two, in which candidates are asked to explain a number of pilot-controller messages. Candidates at level 3 were more likely to realise that they had not understood the message, and were attempting to "guess" the answer. ("*The pilot called for an ambulance but I didn't hear why. Maybe a sick person, maybe a child is ill*".) The high occurrence of **Comparing** suggests that Part Three of the test was working well, with candidates able to compare situations in different environments or over time.



The most commonly occurring function **Advice** reflects the task in Part Two of the test in which candidates are asked to listen to an audio prompt and advise the speaker on an appropriate course of action. This partly explains why giving **Orders** occurs so infrequently. A candidate advising somebody suffering a medical problem might say "*Call for an ambulance on arrival and ask cabin crew to give first aid.*" On the checklists, this would be recorded as **Giving Advice** rather than **Giving an Order**. The same applies to **Requests**, which means to request action from another (not requesting information which is a separate function). The sentence "*Ask cabin crew to give first aid*" does call for action from other people, but again is recorded as **Giving Advice** on the checklist. Requesting or giving **Approval and Permission** is a function which is routinely expressed in Standard Phraseology. TEA deliberately avoids eliciting Standard Phraseology from candidates.

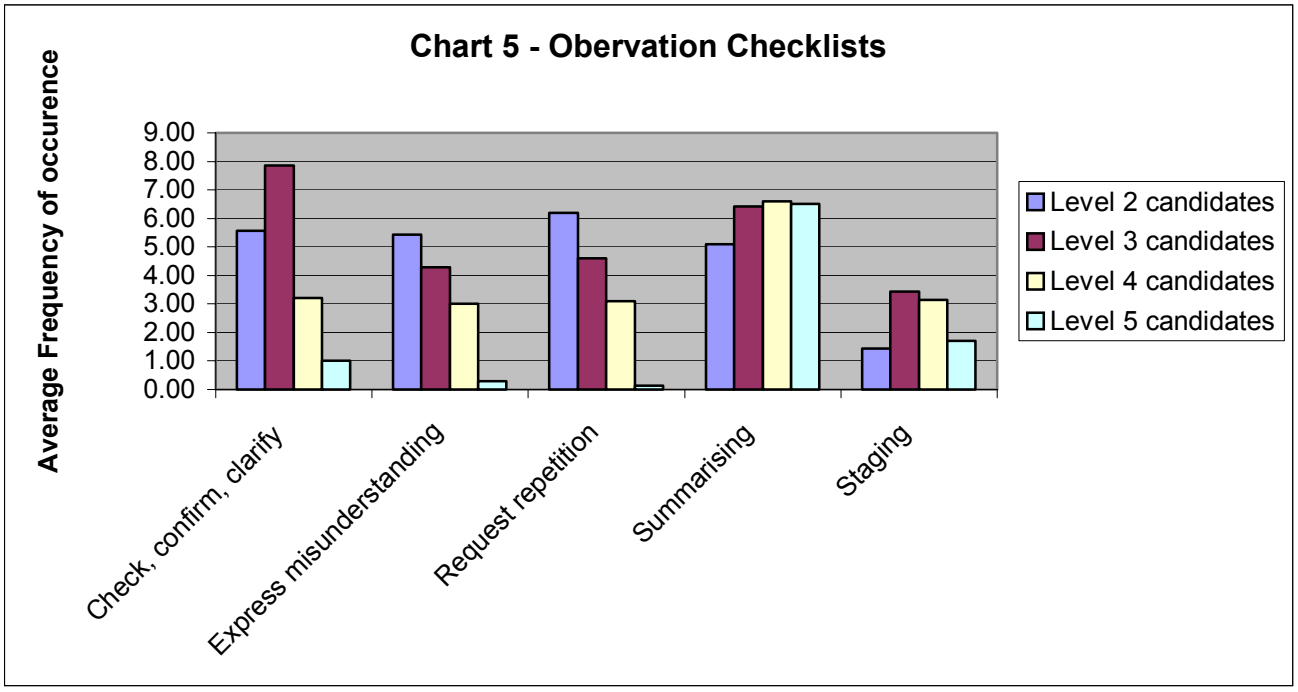


Chart 5: Candidates with weaker comprehension need to check more often and request more repetition than stronger candidates. Lower level candidates also express misunderstanding with greater frequency. The majority of candidates attempted to summarise the messages in Part Two of the test, hence the high frequency of **Summarising** at all levels. However, this does not indicate that candidates gave an accurate summary. **Staging** (setting the context of the next utterance, verbalising the thought process) occurs most at Level 3 as candidates struggle to articulate meaning. Level 3 candidates often repeat the key word of a question to themselves, for instance, before giving a response to the examiner. Higher-level candidates, being able to communicate more immediately, did this much less often.

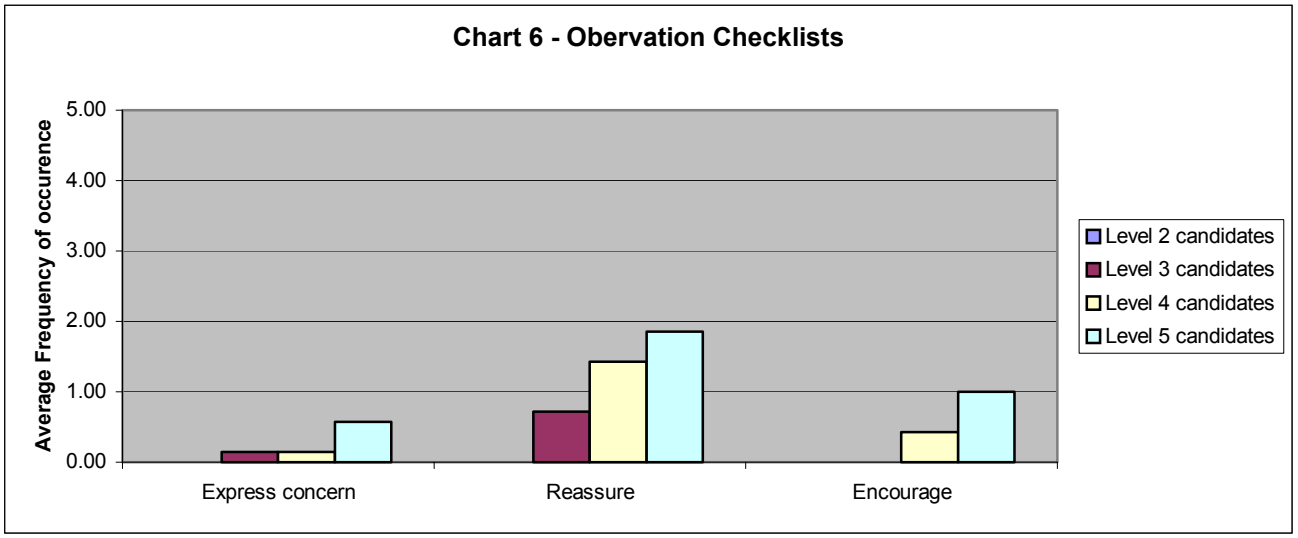


Chart 6: These functions occur with greatest frequency among higher-level candidates. For example, a lower level candidate responding to the audio prompt "*A man has lost his child. Give advice*" might say: *"Go to the Police",* while a higher level candidate might answer:
"I am sorry to hear that. Don't worry. Why don't you contact the Police, I'm sure they can help".

Conclusion

The checklist worked well in its current form and provided useful insights into the performance of the test. The next stage is to apply the checklists to tests in real time (rather than the expensive option of working from transcripts). The results suggest that TEA elicits the functions the test designers intended and Observation Checklists will continue to be used in the trials of new items.

Examiner Training and Reliability

Process of Examiner Training

Examiner training consisted of these stages:

1. Analysis of the ICAO Language Proficiency Rating Scale
2. Group rating (setting the standard)
3. ICAO speech sample calibration
4. Inter-rater standardisation
5. Intra-rater standardisation
6. Ongoing monitoring

The team of senior examiners initially worked with the test design team in studying the ICAO rating scale, the holistic descriptors and ICAO Document 9835 to identify the features of language proficiency that distinguish each level.

The group then applied the rating scale to twenty TEA tests and rated each candidate as a group. The tests were rated aurally, and were also transcribed to allow further analysis, in particular, analysis of lexis and structure. The group reported that whilst the transcripts were helpful, reading them had no significant impact on the marks awarded.

The reliability of the examiners was augmented by involvement in the ICAO rated speech samples project. This was a project to promote international rater calibration. ICAO created a CD featuring candidates at different levels including recordings of the TEA test.

Inter-rater reliability

The group of senior examiners rated 10 candidates of 5 nationalities. Three months later, each examiner was given the tests to rate individually. These marks were then compared to the standard set by the group. The results are presented using the Pearson Correlation Coefficient and the Average Absolute Difference in the marks awarded. The overall mark awarded by an individual examiner is compared to the mark previously awarded by the standardisation group. If the examiner awards the same mark as the standardisation group, the Absolute Difference is 0. On the other hand, if the standardisation group gave a level 2 and the examiner gave a level 4, the Absolute Difference would be 2. Therefore, the closer the Difference is to 0, the more reliable is the examiner's performance. (With the Pearson Correlation Coefficient, 1 indicates perfect correlation, therefore the closer the results are to 1, the more reliable the examiner's marks are.)
The Pearson Correlation Coefficient showed a return of 0.82 and the Average Absolute Difference in marks awarded was 0.2. The group of senior examiners meets on a monthly basis for a standardisation meeting in which tests are rated by the group, with the aid of transcriptions. This facilitates continuous investigations into inter-rater reliability, leading to ongoing examiner training as necessary.

Intra-rater reliability

Examiners were also asked to rate again 5 tests that they had previously rated individually. These tests were selected at random and consisted of different levels of candidates from 4 different nationalities. The examiners re-rated them without reference to the marks they had given before, in order to determine intra-rate reliability. The results again show a high degree of correlation: 0.94 when expressed using the Pearson Correlation Coefficient and 0.13 in the average absolute difference in marks awarded.

New Examiner Training

Prospective TEA examiners undergo a training program, and must certificate before they are allowed to examine. During the training program, the ICAO Language Proficiency Rating Scale and Document 9835 are considered in detail and new examiners are given extensive training with recordings of TEA tests. Prospective examiners then work individually under exam conditions and rate 6 TEA tests. Only if they meet the standard required can they become TEA examiners.

All TEA examiners are subject to ongoing monitoring procedures in order to maintain the reliability of results.

Interlocutor Training and Reliability

In the TEA test, the same person performs the role of both examiner and interlocutor, although the test has been constructed in such a way that these roles can easily be split. Reliability is ensured by using a standardised script. Examiners are trained in the conduct of the test (i.e. performing the role of interlocutor) and must be rated "satisfactory" in this area before they are allowed to examine. The test design team produced a checklist for monitoring the conduct of the exam for this purpose. Having passed this stage of certification examiners are monitored regularly to ensure they continue to conduct the test appropriately.

Concurrent Validity

A study was carried out to compare the results awarded to candidates in the TEA test with assessments of their levels made by teachers who had worked with them. The study was carried out with 47 students from Romania, Kyrgyzstan, Italy, Spain and Brazil. Each group of students were attending a 4-week (100 hours) English for Aviation course to help them achieve ICAO level 4. The teachers working with them were familiar with the ICAO Language Proficiency Requirements and the application of the ICAO Rating Scale, although they were not TEA examiners.

The teachers were asked to provide an assessment of each student's level according to the ICAO rating scale. This assessment was based upon the language demonstrated during the last week of each course. In order to maintain the objectivity of the exam, TEA examiners were not privy to these assessments until after the results of the TEA test had been sent to the candidates.

The assessments of the teachers were compared with the overall score awarded in the TEA test, expressed using the Pearson Correlation Coefficient and the absolute difference in marks. The results demonstrate a very high degree of concurrence between the teachers' assessments and the TEA exam results. When expressed using the Pearson Correlation Coefficient the degree of concurrence is 0.82 and the Average Absolute Difference between the predictions and the marks awarded is 0.26.

	Predicted ICAO Level (by teacher)	Final TEA test score	Absolute difference
Candidate 1	4	4	0
Candidate 2	5	5	0
Candidate 3	4	4	0
Candidate 4	4	5	1
Candidate 5	5	5	0
Candidate 6	4	4	0
Candidate 7	5	4	1
Candidate 8	4	4	0
Candidate 9	3	3	0
Candidate 10	5	5	0
Candidate 11	5	5	0
Candidate 12	5	5	0
Candidate 13	4	4	0
Candidate 14	4	5	1
Candidate 15	6	5	1
Candidate 16	5	5	0
Candidate 17	5	5	0
Candidate 18	4	4	0
Candidate 19	5	6	1
Candidate 20	3	4	1
Candidate 21	5	5	0
Candidate 22	4	4	0
Candidate 23	4	4	0
Candidate 24	5	5	0
Candidate 25	5	5	0
Candidate 26	3	4	1
Candidate 27	5	5	0
Candidate 28	4	4	0
Candidate 29	4	5	1
Candidate 30	4	4	0
Candidate 31	5	5	0
Candidate 32	4	4	0
Candidate 33	4	4	0
Candidate 34	4	5	1
Candidate 35	4	4	0
Candidate 36	2	3	1
Candidate 37	4	4	0
Candidate 38	4	4	0
Candidate 39	3	4	1
Candidate 40	3	4	1
Candidate 41	4	4	0
Candidate 42	3	3	0

Candidate 43	3	3	0
Candidate 44	3	3	0
Candidate 45	3	3	0
Candidate 46	3	3	0
Candidate 47	4	4	0

Stakeholder Feedback

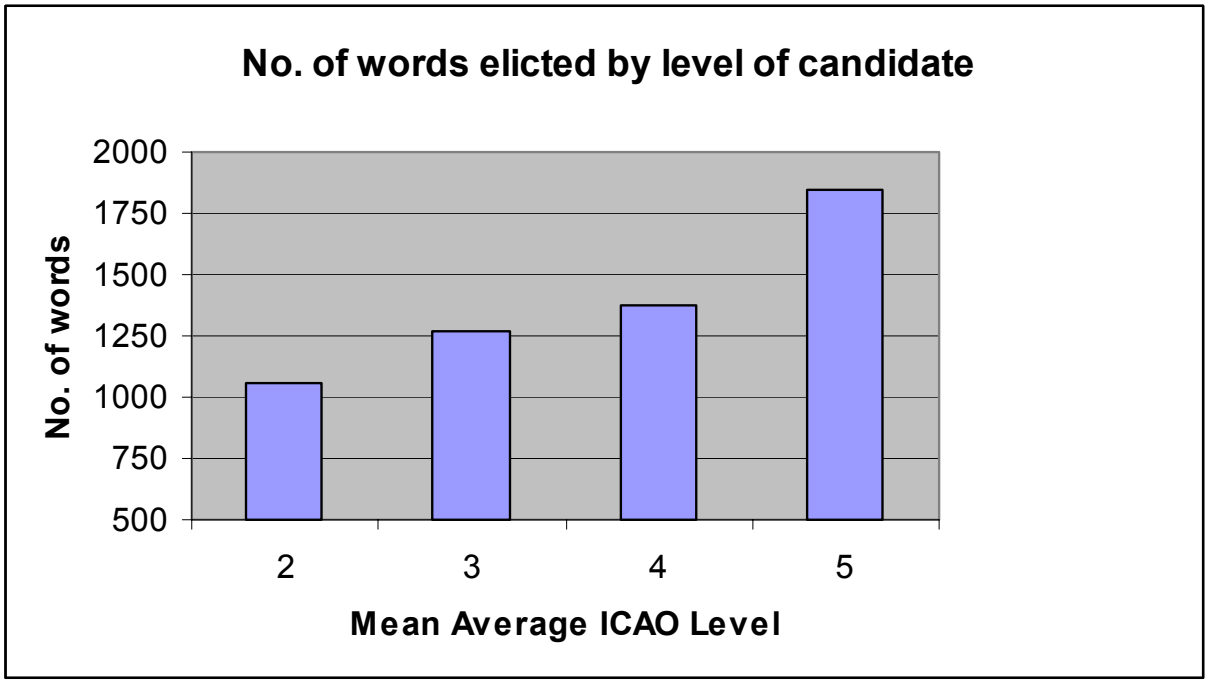
The Concurrent Validity study was augmented with feedback from other stakeholders who were asked to give their assessment of TEA's performance. A questionnaire was sent to aviation training institutes which had used TEA to test their personnel. We asked these institutes whether the levels awarded to their students matched the expectations of teachers who had worked with them, and whether they felt that TEA was appropriate for benchmarking and licensing purposes. Those questioned reported that the overall levels candidates achieved were generally in line with their expectations. Whilst it is encouraging that stakeholders feel that TEA is reliable and valid, we recognise that this form of research has its limitations. Although we know that the results of the test met the institutes' general expectations, there is no data showing the predicted level of each candidate. Due to the number of candidates involved it was not possible to conduct this study with the same precision as is possible with a smaller sample.

Word Count

The transcripts of 36 candidates were analysed in order to determine the average number of words spoken by the examiner and the average number of words spoken by the candidates (broken down by level of candidate.

The results were:

Average number of words spoken by a TEA examiner – 486.



	Average ICAO Level	Word Count
Candidate 1	2	1049
Candidate 2	2	913
Candidate 3	2	1113
Candidate 4	2	943
Candidate 5	2	1205
Candidate 6	2	1168
Candidate 7	2	1385
Candidate 8	2	682
Candidate 9	3	992
Candidate 10	3	1133
Candidate 11	3	1385
Candidate 12	3	1352
Candidate 13	3	1254
Candidate 14	3	1055
Candidate 15	3	1364
Candidate 16	3	1171
Candidate 17	3	1214
Candidate 18	3	1384
Candidate 19	3	1730
Candidate 20	3	1722
Candidate 21	3	2094
Candidate 22	3	966
Candidate 23	4	1467
Candidate 24	4	1371
Candidate 25	4	1795
Candidate 26	4	1421
Candidate 27	4	1335
Candidate 28	4	1480
Candidate 29	4	1270
Candidate 30	4	1021
Candidate 31	4	1204
Candidate 32	5	2110
Candidate 33	5	1985
Candidate 34	5	1851
Candidate 35	5	1858
Candidate 36	5	1926

Test Security

High-stakes testing

Security is a major issue in the high-stakes environment of aviation testing. There are a number of potential threats to the security of any test. In particular:

- Materials being leaked
- Impostors taking the test
- Fraudulent certificates

TEA security features

TEA has a number of security features to reduce these risks. Candidate applications include biographical data, photos, signatures and passport numbers which can be checked prior to the exam.

To reduce the risk of materials being compromised, there are multiple versions of the test (to make it more difficult for candidates to prepare answers). All test materials are stored securely at test centres and inventories are kept updated. Candidates are not permitted to take any electrical equipment, including mobile phones, into the test, to reduce the risk of the test being recorded.

Test day security

Immediately before a candidate takes the test, the examiner checks the passport and biographical data and photographs the candidate. This photograph includes the date and time and appears on the certificate. (An impostor would not only have to fake identification but also have to change their appearance.)

Certificates

Certificates are produced centrally and contain a number of security features to reduce the risk of forgeries: Candidate photo, a unique certificate number which can only be verified by Mayflower College, micro-printing, tinted colours, a verification stamp, and the signature of the Director of Testing.

To date, there have been no recorded incidents of security breaches.

Bibliography

Amendment to the Note in the Appendix to Annex 1 — Personnel Licensing. ICAO. 13 June 2006.
Converting an Observation Checklist for use with the IELTS Speaking Tests. Lindsay Brooks. Cambridge ESOL Research Notes 11 February 2003.
Developing Observation Checklists for Speaking Tests. N Saville and B O'Sullivan. Cambridge ESOL Research Notes 3. November 2000.
Does Washback Exist? . J. C Alderson and D Wall. Applied Linguistics 14.
Evidence-based validation: what makes a test 'specific'. B O'Sullivan. ALTE Cardiff November 2005.
Examining Concept Maps as an Assessment Tool. M. Araceli Ruiz-Primo. Stanford University.
Fundamental Considerations in Language Testing. L F. Bachman. OUP. 1990.
Heightened awareness of communication pitfalls can benefit safety. B. Day. ICAO Secretariat.
ICAO Language Proficiency Requirements. E. Matthews. ICAO.
Insights into the FCE Speaking Test. Y. Lu. Cambridge ESOL Research Notes 11.
Introspection in Second Language Research. Faerch and Kasper. Clevedon. Multilingual Matters. 1987.
Issues in Speaking Assessment Research. L. Taylor. Cambridge ESOL Research Notes 1.
Language Test Construction and Evaluation. J. C Alderson, C Clapham and D Wall. CUP. 1995.
Language training and testing in aviation need to focus on job-specific competencies. J. Mell, Ecole Nationale De L'Aviation Civile, France
New provisions for English language proficiency are expected to improve aviation safety. E. Matthews. ICAO Secretariat.
On Taking Tests: *What the Students Report.* A.D. Cohen Language Testing 1. 1984,
On the validity of cognitive interpretations of scores from concept-mapping techniques. M. Araceli Ruiz-Primo, R. Shavelson, M. Li, S. Schultz. Educational Assessment 7.
Principles and Practice in Test Development. L Taylor. Cambridge Research Notes 3.
Review of Recent Validation Studies. Cambridge Research Notes 9.
Revising the IELTS Speaking Test. L. Taylor. Cambridge ESOL Research Notes 5.
Standards for Safety – The Language Barrier. El-Kadur.
Standards for Test Development in Aviation: from placement to proficiency. J. Mell. Ecole Nationale del'aviation civile (ENAC).
Test Validation and Cognitive Psychology: *Some Methodological Considerations*. Language Testing 3.
The ALTE Code of Practice. Association of Language Testers in Europe. Cambridge 1994.
The Critical Components of Aviation English. M. Mitsutomi, K. O'Brien.
The Development of a Set of Assessment Criteria for Speaking Tests. A. ffrench. Cambridge ESOL Research Notes 13.
Using Observation Checklists to Validate Speaking Tests. Barry O'Sullivan, Cyril J. Weir and N Saville. Language Testing 2002 19.