Report 11 - Examiner Training & Assessment using TEA

This report will firstly outline the approach to Examiner Training before describing the method of assessment in TEA. For more detailed analysis of three candidate performances, see *Report 12 – Example Criteria in TEA Assessment - 3 Annotated Test Transcripts*.

A – Initial TEA Examiner Training & Certification

Prospective TEA Examiner Requirements:

- A background in either English language teaching and/or operational aviation experience
- Minimum IELTS 7.0 Overall, including 7.0 in Speaking & Listening (guided by EANPG, Appendix J & K (30 Nov 2006)) or equivalent (a TEA certificate at Level 6)
- Considered capable of the required level of professionalism and of adherence to ILTA Code of Ethics.

TEA Examiner candidates undergo a face-to-face 5-day training program, and must be certified before they are allowed to examine. During the training program, the ICAO documentation is considered in detail and candidates are given extensive rater training with recordings of TEA tests. Candidates then work individually under exam conditions and rate 6 TEA tests. Only if they meet the standard required can they become TEA Examiners.

TEA Examiner training consists of the following stages:

1. Analysis of the ICAO Language Proficiency Requirements (*9835*), the Rating Scale & the ILTA Code of Ethics
2. Group Rating (setting the standard)
3. Guided Rating Practice
4. Rating Certification
5. Intensive Interlocution Training
6. Interlocution performance demonstration – rated 'satisfactory' before allowed to conduct live tests
7. TEA Administration & Security

To become TEA Examiners*, candidates must pass both the rating and interlocution aspects: course attendance or completion in no way guarantees TEA Examiner status.

[*In some organisations, there is a requirement for test performances to be externally rated. In such cases, TEA Examiners are interlocutors only and tests are rated remotely by TEA Monitoring Examiners. This system is in accordance with 9835 (6.3.4.1)]

Interlocutors are trained to elicit an appropriate language sample while consistently delivering the test according to stated procedures. A reminder of the procedures is also printed at the beginning of every TEA Examiner Handbook, titled *TEA Test Interlocution Guide* (see APPENDIX A). During training, checklists are used by both instructor and trainee to monitor performance development. A

sample *TEA Interlocutor Self-Assessment Form 2010* can be found in APPENDIX B. The *ILTA Code of Ethics* is also studied in detail to instil principles of best practice in examiners.

B – TEA Assessment

(*See APPENDIX C for a rationale of TEA's stance on rating systems and 'primary' raters.*)

The driving force of assessment in the field of testing plain English for aviation is the ICAO Descriptors. In the text below, italicised text (in blue) is wording taken from the ICAO Descriptors.

TEA is a profile-assessed test of English designed to elicit the language required for the application of the ICAO proficiency descriptors. Global marking is not considered appropriate for this purpose, as the six areas of proficiency must be rated independently (from *9835*, 4.5.5. "(d) the final rating is not the average or aggregate of the ratings in each of the six ICAO language proficiency skills but the lowest of these six ratings."

The test is in 3 sections, each designed to offer the opportunity to demonstrate language proficiency and appropriacy across the full range of candidate ability in *'common, concrete, or work related topics'* (ICAO bands 2 and above) and when *'confronted with a linguistic or situational complication or an unexpected turn of events'* (ICAO bands 4/5/6).

Part 1
This is designed to act initially as a warmer moving into comprehension and description/explanation/expansion of *'common, concrete, or work related topics'*. The focus for the rater in this section is the candidate's ability to comprehend the interlocutor's prompts and provide *'immediate, appropriate and informative'* responses.

Part 2A
The focus in part 2A is on comprehension of *'common, concrete, or work related topics'* delivered in a variety of international accents, *'sufficiently intelligible for a community of international users'*. There is an opportunity for higher level candidates to demonstrate the ability to successfully paraphrase and use idiomatic language.

Parts 2B & 2C
The focus in parts 2B &2C is on the candidate's ability to interact with prompts delivered in a variety of accents *'sufficiently intelligible for a community of international users'*. The prompts are in an aviation context but with a *'linguistic or situational complication or, unexpected turn of events'*. Therefore parts 2B & 2C offer a higher-level candidate (level 4 but especially levels 5 & 6) the opportunity to demonstrate an ability to comprehend and interact in this type of situation by asking questions and giving advice, two functions of speech listed in 9835. The range of vocabulary needed for this part of the test is wider than that needed for parts 1 and 2A and may require the candidate to paraphrase when lacking vocabulary in *'unusual or unexpected circumstances'*.

Part 3
Describing a situation in an aviation context in the form of a set of pictures enables candidates of all but the lowest levels to demonstrate their proficiency in Fluency, Pronunciation, Vocabulary and Structure. A level 2 candidate may offer some *'isolated words and memorised phrases'*. A level 3 candidate may respond using simple vocabulary and *'basic grammatical structures'* which are not always well controlled resulting in frequent

interference with meaning. A level 4 candidate would offer a slightly wider, more accurate range of vocabulary using *'basic grammatical structures'* that are *'usually well controlled'*. A level 5/6 user will accurately and appropriately use a wide range of vocabulary which is sometimes idiomatic, and make use of a range of both *'complex grammatical structures'* and consistently well controlled basic structures. The candidate may also use a range of language functions beyond simple descriptive functions. These may include for example; justification, speculation, and expressing concern.

The candidate then enters into a discussion with the interlocutor based on the topic of the second set of pictures. Although in an aviation context, part 3 prompts are on a range of less familiar topics thus allowing higher level candidates, (levels 5/6), the opportunity to demonstrate a much wider range and proficiency of Vocabulary, Structure, Fluency and Interactions. A level 4 candidate may offer simple responses with the need to use clarification strategies and paraphrase. A candidate at level 3 or below may be unable to offer appropriate responses or at best will produce very simple responses in this final part of the test.

In which parts of the test are each of the profiles assessed?

Pronunciation is assessed throughout the test.

Structure is assessed throughout the test with higher level candidates, 5/6, being able to demonstrate both basic and complex grammatical structures in part 3.

Vocabulary is assessed throughout the test. Higher level candidates (4 or especially 5/6) should demonstrate a wide range of vocabulary especially in parts 2B, 2C and 3.

Fluency is assessed throughout the test although the opportunity for the candidate to speak at length is offered in part 3.

Comprehension of *'common, concrete, or work related topics'* is assessed in parts 1 and 2A and comprehension of a *'linguistic or situational complication or an unexpected turn of events'* in parts 2B, 2C and 3. Part 2 is designed to test comprehension of a range of international accents *'sufficiently intelligible for a community of international users'*.

Interactions are assessed in parts 1, 2B, 2C and 3. The assessment in part 1 is aimed at all levels. Parts 2B & 2C are aimed in particular at testing the candidate's ability to respond adequately and appropriately with an unexpected turn of events. In addition, parts 2B &2C require candidates to ask questions and give advice, two of the functions listed in *9835*.

Candidates are awarded a score for each profile, the lowest score determining their overall score (as decreed by *9835*). Raters are trained to identify features of a linguistic performance and award scores based on the most appropriate description in each profile. Since both productive and receptive skills are being tested, it is necessary to focus on the assessment of Comprehension independently. The following text is taken from *Rating of Comprehension 2010* (in which 2010 refers to TEA <u>Version</u> 2010):

Overview

The focus in part 2A is on comprehension of *'common, concrete, or work related topics'* delivered in a variety of international accents, *'sufficiently intelligible for a community of international users'*. There is an opportunity for higher level candidates to demonstrate the ability to successfully paraphrase and use idiomatic language.

Comprehension of *'common, concrete, or work related topics'* is assessed in parts 1 and 2A and comprehension of a *'linguistic or situational complication or an unexpected turn of events'* in parts 2B, 2C and 3. Parts 2A, 2B and 2C are designed to test comprehension of a range of international accents *'sufficiently intelligible for a community of international users'*.

Application

The application of the above is a combination of a numerical score (part 2A) and the application of a scale of proficiency by trained raters (parts 1, 2B, 2C and 3).
The numerical scoring of part 2A sets a minimum requirement but in order to achieve level 4,5 or 6 the candidate must demonstrate adequate comprehension of part 1 (*'common, concrete, or work related topics'*) and for level 5 or 6, mostly or consistently accurate comprehension of parts 2B, 2C and 3 (*'linguistic or situational complication or an unexpected turn of events'*) respectively.

Numerical Scoring 2010

Part 2A comprises 10 short recordings in which either a pilot or a controller communicates a message about a non-routine situation. **1 point is awarded for each correct response**. **There is a maximum of 10 points for 2A**. A response is considered correct when a candidate demonstrates a complete understanding of the *sense* of the whole message, rather than simple, disconnected references to parts of the message. **Half-points cannot be awarded**. **The information can be relayed using exact wording or paraphrase, and should be unambiguous**. If paraphrase is used, marks can only be awarded if the essence of the original message is maintained. Extra ideas from the candidate should not distort the meaning of the original message. Since candidates are told that all the information is important, that they can hear recordings a second time to clarify their understanding, and are asked "*what was the message?*", it is not unreasonable to expect candidates at level 4 and above to accurately relay the sense of the message, including the important details, most of the time.

Raters must consider various aspects of the candidate's response to each situation.
For example, **does the candidate confuse**:

- *who the speaker is? (Is it a pilot or a controller?)*
- *what the speaker wants or is saying?*
  (Is it a request, a confirmation, a question, a command?)
- *the tense (where tense affects meaning)?*
- *whether the situation is resolved, or needs further action?*
- *the urgency of the situation?*
- *the key elements of each message?*

Where it is not clear to the rater that the candidate has understood all aspects of the recording, the candidate is awarded 0 points. **Half-points cannot be awarded**.

The following examples are intended to illustrate incorrect and correct responses to sections of the 2A recordings:

	Points Awarded

Recording 1: Do you have fire, crash and rescue services?

"They're requesting fire, crash and rescue services."	*0*
"They want to know if airport has fire, crash and rescue services."	*1*

Recording 2: Our sick passenger is stable now.

"The sick passenger was stable."	*0*
"The passenger is better now."	*1*

Recording 3: We are running low on fuel.

"They have no fuel."	*0*
"They are short of fuel."	*1*

Recording 4: Cancel the request for an ambulance.

"They need an ambulance."	*0*
"They don't need an ambulance."	*1*

Recording 5: We need some medical assistance because a passenger is sick.

"They need some help for the passenger."	*0*
"They need some medical help for the sick passenger"	*1*

Repetitions
For level 4 comprehension,
       there is no limit to the number of repetitions a candidate can ask for.
For level 5 comprehension,
       the candidate should not require regular repetition during the test.
For level 6 comprehension,
       the candidate should require no more than a minimal number of repetitions during the test.

The numerical scoring system in 2A creates strict *ceilings* (cut-off points) which, in conjunction with the application of the proficiency scale, help raters to assign a level to each comprehension performance:

Part 1 Comprehension	Points in 2A	Parts 2B, 2C & Part 3 Comprehension	OVERALL Comprehension Band
adequate*	9 or 10	consistently accurate	6
adequate*	9 or 10	mostly accurate	5
adequate*	8	consistently or mostly accurate	5
adequate*	8	slow or inadequate	4
adequate*	7	consistently or mostly accurate	5
adequate*	7	slow or inadequate	4
adequate*	6	consistently or mostly accurate	4
adequate*	6	slow or inadequate	4
inadequate	6	not applicable	3
adequate*	5	consistently or mostly accurate	4
adequate*	5	slow or inadequate	4
inadequate	5	not applicable	3
not applicable	1 to 4	not applicable	3
not applicable	0	not applicable	2
*adequate = mostly accurate			

This assessment system was designed to fall in line with the ICAO Descriptors. For example, candidates can only get level 4 for Comprehension by demonstrating adequate comprehension in Part 1 and a minimum of 5 situations in Part 2A i.e. "mostly accurate (comprehension) on common, concrete, and work related topics when the accent or variety used is sufficiently intelligible for an international community of users"; furthermore, a Comprehension Level 4 candidate may struggle to understand elements of Parts 2B, 2C & 3 i.e. "when the speaker is confronted with a linguistic or situational complication or an unexpected turn of events, comprehension may be slower or require clarification strategies"; additionally, there is an allowance for candidates at Comprehension Level 4 to ask for repetition to account for "comprehension may be slower or require clarification strategies".

Through the adoption of this system in April 2010, plus improved training and standardisation, rater subjectivity has been further minimised and candidates at all levels can be assessed in line with the ICAO Descriptors. For more information about rater standardisation, see **Report 13 – TEA Examiners & Rater Reliability**.

Concurrent Validity

Concurrent validity is a measure of how well a test scores correlate with a previously validated measure. In the context of plain English for aviation, this has been impossible since no measures of like-for-like assessment in this testing context have been validated. In that respect, TEA fully recognises the limitations of this type of study.

Research was conducted into the validity of test scores of candidates who had participated in Aviation English courses at Mayflower College, Plymouth during the calendar year of 2010. 201

tests were conducted with students from Azerbaijan, Brazil, Bulgaria, Colombia, Estonia, Holland, Italy, Kyrgyzstan, Morocco, Poland, Russia, Spain, Ukraine and Uruguay.

The study involved comparing TEA Overall Scores with previously given assessments of their levels made by teachers who had worked with them. The 4 teachers were familiar with the ICAO Language Proficiency Requirements and the application of the ICAO Rating Scale, although they were not TEA Examiners. Their assessment was based upon language proficiency demonstrated during the last week of each course. In order to maintain rating objectivity, TEA Examiners were not shown the teacher-assessments.

The assessments of the teachers were compared with the candidates' Overall Score awarded in the TEA test, expressed using the Pearson Correlation Coefficient and the Absolute Difference between scores.

PEARSON	0.923431613
Total number of tests	201
Total of absolute differences	28
Average absolute difference between scores awarded	0.14

The results demonstrate a very high degree of correlation between the teachers' assessments and the TEA scores. Of 201 tests, there was disagreement between predicted and actual score in only 28 cases (and there were no differences of 2 band scores). When expressed using the Pearson Correlation Coefficient, the degree of concurrence is 0.92 and the Average Absolute Difference between the predictions and the Overall Scores awarded is 0.14. It seemed that TEA was able to accurately reflect 'true' proficiency most of the time.

APPENDIX A

TEA Test Interlocution Guide

Here are some reminders of good interlocution practice when delivering TEA. Interlocutors are constantly monitored and failure to adhere to these guidelines may result in a requirement to re-train and re-certify as a TEA Examiner.

It is essential to go through the test materials carefully in order to be completely familiar with them and to handle materials naturally and smoothly. Instructions should be delivered naturally, clearly and audibly. **If the candidate expresses a lack of understanding of the instructions at any time, you should paraphrase the instructions briefly for them.**

It is also imperative you familiarise yourself with any technical terms used in the test. Clearly, it is not possible to predict everything that a candidate might say. If a candidate uses technical terms which you do not know, ask the candidate to explain or, if this is inappropriate, check an aviation dictionary (after the test is completed). Alternatively, contact your TEA Administrator for advice.

Remember to vary sets and tasks as much as possible, using tasks within each section randomly - don't select or avoid. Make the transitions between parts clear to the candidates. Accurately record the version numbers (highlighted in blue) used during the test.

Introduction

Record the introduction before the candidate enters the test room. Check the candidate's identity document carefully when they are in the room.

Part 1

Ask the question set relevant to the candidate – Commercial Pilots and Controllers, Private Pilots, Ab-Initio Pilots or Student Controllers.

Ask the questions in order. Only omit a question if a candidate has answered it in an extended previous answer.

Do not deviate from the script at all (except for glossing - see below). Only use the extra prompts in brackets when the closed question elicits a one-word or very short response.

You may gloss (briefly paraphrase) individual words if the candidate asks for clarification.

If you suspect an answer is memorised, interrupt with *"Thank you"* *and ask the next question.*

Adhere to the timing for this section: 5 to 6 minutes.

Clearly introduce the next section of the test.

Part 2A

Note-taking is not permitted.

Give the initial instructions clearly – stressing the key words.

Say "What was the message?" after every recording. This breaks the memory loop, forcing language to be processed beyond simple repetition.

Do not play a recording twice unless the candidate specifically asks for repetition.

Allow the candidate 10 seconds to begin their response and a maximum of 20 seconds to respond. If they have already responded but would like to hear the recording again, asking for repetition within the 20 seconds, play the recording a second time.

You can move onto the next recording before the 20 seconds is up only if it is clear that the candidate has nothing more to add.

Parts 2B & 2C

Give the initial instructions clearly – stressing the key words.

Give the appropriate prompt – either "Ask Questions" or "Give Advice" – after each recording.

If the candidate does not follow the instructions, politely but firmly repeat the prompt. If the candidate does not understand, say "OK, thank you", and continue.
If the candidate expresses a lack of understanding of the instructions, quickly paraphrase the instructions for them.

Allow the candidate 10 seconds to begin their response and a maximum of 20 seconds to respond. If they have already responded but would like to hear the recording again, asking for repetition within the 20 seconds, play the recording a second time.

You can move onto the next recording before the 20 seconds is up only if it is clear that the candidate has nothing more to add.

Clearly introduce the next section of the test.

Part 3

Give the initial instructions clearly – stressing the key words.

Before describing the pictures, try to ensure the candidate uses their preparation time. Break eye contact, pretend to make a note of something, etc so that they don't start talking too early.

Do not allow the candidate to start talking about the pictures before you have given the full instructions.

Do not allow more than 60 seconds for the descriptions.

Clearly introduce the discussion section of the test.

There are 3 potential sub-topic question sets and the candidate should be encouraged to explore the topics as fully as possible. Ask suitable follow-up questions in a natural way to encourage the candidate to justify, clarify or explain their opinions.

If the candidate does not understand a question, you are free to paraphrase it in this section.

Introduce each sub-topic clearly. To lead into a second topic, use a bridging phrase such as:

"Thank you. Now let's talk about …. [sub-topic title]".

Begin with any question set, asking each question in that set in order before moving onto another sub-topic. If a candidate answers the next question in an extended previous answer, skip the question.

Extend the candidate's English as much as possible. Keep to the topic, but ask them to expand on opinions (e.g. "Why do you think that?"). If a candidate gives a one word answer, attempt to extend their English by asking "Could you tell me more?".

Adhere to the timing for the discussion: 4 to 5 minutes.

APPENDIX B

TEA Interlocutor Self-Assessment Form 2010

Having recorded your practice TEA Test, please listen back to the recording and use this self-assessment form to assess your interlocution performance. The statements marked with * may or may not apply to your practice test.

PRACTICE TEA TEST		Examiner Checklist	Yes/No
		As the examiner, I…	**(please tick)**
Pre-Test	1	prepared well and knew my materials well.	
Introduction	2	recorded the introduction before the candidate entered the test room.	
	3	checked the candidate's identification	
Part 1	4	asked a question set relevant to the candidate.	
	5	did not deviate from the scripted questions.	
	6	asked the questions in order.	
	7	skipped a question that had already been answered in an extended previous answer.*	
	8	successfully glossed a word that the candidate didn't understand.*	
	9	used extra prompts only when appropriate.	
	10	stuck to the timing for this section: 5 – 6 minutes.	
Part 2	10	clarified the instructions to the candidate when asked.*	
	11	allowed the candidate a maximum of 10 seconds to begin speaking, and a total of 20 seconds to respond.	
	12	did not help a candidate, either verbally or non-verbally, to understand the recordings.	
	13	did not give positive or negative responses to the candidate after their response.	
	14	did not play a recording twice unless the candidate specifically asked for repetition.*	
	15	stated "OK, I'll play it once more" when repeating a recording.*	
	16	did not play a recording more than twice.	
Part 2A	17	said *"What was the message?"* after every recording.	
Part 2B	18	ensured the candidate responded appropriately by saying *"No, ask the speaker questions"* if necessary.	

Part 2C	19	ensured the candidate responded appropriately by saying *"No, give advice to the speaker"* if necessary.	
Part 3 Pictures	20	gave the initial instructions clearly, emphasizing key words.	
	21	allowed the candidate 15 seconds to study the pictures and prepare to speak.	
	22	did not allow more than 60 seconds for the picture description.	
	23	took the pictures back after 60 seconds.	
Part 3 Discussion	24	introduced each sub-topic clearly.	
	25	asked the questions in order.	
	26	skipped a question that had already been answered in an extended previous answer.*	
	27	successfully rephrased/explained a question that the candidate hadn't understood.*	
	28	extended the candidate's English as much as possible (while keeping to the topic) by asking them to expand on opinions or one-word answers.	
	29	used an appropriate bridging phrase to lead into a new sub-topic.	
	30	adhered to the timing for the discussion: 4 to 5 minutes.	

TEA Examiners must demonstrate satisfactory compliance with the aspects of interlocution listed above. **For the sake of test reliability, we can only certify TEA Examiners who are consistently able to follow the interlocution guidelines.** The same standards must also be met by our current examiners.

APPENDIX C

The majority of TEA tests are rated by one 'primary rater'. A strict monitoring policy is then applied to check examiner standards and a third rater is employed where a disagreement emerges. In a limited number of cases, the national aviation authority has decreed that two primary raters are necessary. In such cases, the raters discuss their assessments and agree what scores are to be entered on the TEA Database against one of the rater's names.

9835 states:

> "*Ideally, an aviation language test will have two primary raters — one language expert and one operational expert — and a third rater who can resolve differences between the two primary raters' opinions. For example, there could be a situation where the primary raters agree that in five of the six skill areas a test-taker demonstrates Level 4 proficiency; however, the first rater assigns the test-taker a score of 3 on pronunciation (thereby making the test-taker's overall language proficiency level "3") and the second rater assigns the test-taker a "4" for pronunciation. A third rater would make a final determination for that skill area and, in doing so, would determine the overall score for that test-taker. A third rater would likely be involved in the process only in cases in which a test-taker may obtain an overall rating of 3 or 4, since the difference between these two levels is the most critical distinction for ICAO language proficiency licensing testing.*" (6.3.4.2)

TEA believes that although the two primary rater format is theoretically desirable, the practical difficulties it can create make it a fundamentally difficult system to impose for the following reasons:

- *What is a disagreement?* If rater 1 scores a candidate 333444 and rater 2 scores him 444333, they have arrived at the same Overall Score while simultaneously being in 100% disagreement. It is clear that it is not only the Overall Score that matters to operational personnel and institutions – the ICAO Rating Scale demands that candidates are assessed using profile marking across 6 'sub-skills'. Negative washback and a negative effect on training plans would result from such unreliability. In many cases, 2 examiners will not agree on all 6 profile scores and so a third rater would be required most of the time.
- *Whose decision is final?* In a 2 rater system, assessment could be compromised by unreliable factors such as status and experience: a power relation or other role conflict can lead to unreliable results.
- *Is it fair?* Candidates who have to face two examiners simultaneously are put under greater pressure than those who face only one.
- *Is there time?* In terms of test administration, a 2 rater system is more difficult to organise. If 2 raters are to be present in every test room, the organisational difficulties are clear. If one or both raters are to rate remotely, organising the assessment is more complicated and time-consuming. A delay in certification would result.
- *Are customers prepared to pay?* Although the term 'affordable safety' appears crass in a high-stakes testing context, it cannot be ignored. The majority of customers have to balance the need for quality testing against budget constraints.
- *Is there a conflict of skills? 9835* also states that "*the involvement of operational experts such as pilots, controllers and flight instructors or examiners in the rating process will add operational integrity to the process. Operationally experienced raters can also assist by making informed judgements from an operational perspective on such aspects of language use as conciseness (exactness and brevity) in speech and intelligibility of accents and*

dialects that are acceptable to the aeronautical community" (6.3.8.9). TEA believes that there is an inconsistency between the guideline of 6.3.8.9 and the ICAO Rating Scale. While having operational experts train and certify as examiners is potentially beneficial to the integrity of the process, the language elicited in TEA is to be rated by a linguistic expert – a person with proven language proficiency and proven ability to apply the Descriptors consistently and accurately. Given that the Rating Scale does not mention 'conciseness' as a positive feature of language performance but, in fact, rewards speaking '*at length*', there is a clear conflict within the guidelines. Furthermore, to speculate that an operational (but linguistically untrained) rater may be more adept at assessing international intelligibility appears awkward to justify. TEA does not believe that an operational expert would necessarily have had more exposure to 'accents and dialects' than a linguistic exert.

Other high stakes tests (e.g. IELTS, from which scores are used for immigration and course/job entry purposes) use a one-rater system but with a less vigilant monitoring policy (tests are only double-marked when a 'jagged' profile emerges i.e. when 2 or more scores on sub-tests are not in line).

In summary, TEA believes that through good training, strict certification/re-certification, regular standardisation and the implementation of an effective monitoring policy, a one primary rater approach is the most valuable, efficient system for this context.