Report 08 – Item Development & Version Content

This report will outline the process of developing, trialling and revising items for TEA, describing each stage for each task of the test (see ***Report 03 - Description of Tasks & Instructions*** for detailed information about tasks and rubrics). Also, the management of version content – how different TEA versions (or *forms*) are balanced in terms of content – will be described.

The report is ordered into the following sections:

A – TEA Handbooks

B – Item Writers

C – Item Development for each task:
- Part 1
- Part 2A
- Parts 2B & 2C
- Part 3 Pictures
- Part 3 Discussion

Within each section:
- Item-writing Framework
- Item Trialling & Item Revision

D – Version Content

A – TEA Handbooks

TEA Handbooks are developed every 6 months and are written, trialled and produced in pairs (A & B, C & D, E & F, etc). Pairs of new handbooks are sent securely to the busiest TEA Centres and previously-used handbooks are withdrawn. Smaller TEA Centres may receive handbooks in stages depending on demand. Every effort is made to give centres fresh test materials while balancing both test security and the circulation of current materials. Withdrawn test materials may be recycled at a later stage although handbooks are never re-produced in their entirety.

Notice of live and withdrawn test materials is organised through the Materials Development Manager in co-ordination with the TEA Centres Manager, the Central TEA Administrator and the individual Administrators at TEA Centres around the world.

Each Handbook contains:
- A Guide to Interlocution
- Instructions to examiners for introducing the test
- Test Rubric (including test items) for each part of the test
- The ICAO Descriptors

- 6 picture sets
- 4 CDs.

B – Item Writers

Item writing is conducted in the form of a collaborative process. Item writers are drawn from both a wide range of linguistic and operational backgrounds, and a variety of nationalities. First-draft writers are native speakers of English with qualifications and experience in Linguistics, Language Testing and/or Language Teaching. Second-draft writers are high-level English users with either 5 years of operational experience, or post-graduate qualifications.

For example, the test items for Part 2A – the short, non-routine messages from both pilots and controllers delivered semi-directly to test candidates – are first written in line with the test-writing framework in the UK, before experienced operational pilots and controllers check the items for both operational authenticity and technical accuracy. During the last draft of item writing, proof-readers were drawn from contacts in Azerbaijan, Bulgaria, Colombia, Italy, Spain, Sweden, and the USA. Since the messages are written in plain English and aviation phraseology is avoided, item re-drafting is frequent and an open channel of communication between the item writers and the operational proof-readers is vital.

All item writers are given the following to enable them to complete their work with accuracy and relevance on the given test task:

- A copy of the test specifications and a description of the language competences the task is intending to engage
- A framework for the task including the language functions to be elicited
- Samples of previous task items
- Guiding language in the form of appropriate domains and vocabulary*.

TEA developed a *Word List* appropriate to the context of plain English by combining the 'Globish 1500' with a selection of aviation-specific vocabulary such as *aerodrome*, *bird strike*, *collision*, *divert*, etc. Globish is a subset of the English language formalized by Jean-Paul Nerriere which uses a subset of standard English grammar, and a list of 1500 English words. The list has come to be recognised as the common ground that non-native English speakers adopt in the context of international business.

C – Item Development

Fulcher and Davidson describe how test development should occur:

> "Tests should be built to the best of an organisation's ability at the time that we first create them, and then, as and when matters change and the test needs refitting, the organisation should refit it. A test should not remain in place simply because it is in place. Stasis is both a blessing and a curse, but its mixed nature can be moderated if all parties are willing to talk about the

test – and such dialogue should happen at a higher, more productive level of conversation: it should be discussion about the specs and not only about the test." (2010: 60)

The last 'refit' of TEA was undertaken with consultation from all parties and the consequential changes became TEA Version 2010, the item development for which is described for each task in turn below.

Part 1

Item-writing Framework

To write 6 question sets (for 2 handbooks), the following guidelines should be adhered to:

- Each set consists of 7 questions related to the candidate's aviation role in each set. Their role – Operational Pilot or ATC, Private Pilot or Student Controller – must be considered carefully in question writing since there are differences, and therefore limitations for item-writing, between each.

- To relax the candidate and to focus on 'common, concrete, and work related' topics, 5 initial questions are primarily focussed on the simpler language functions, listed in the left-hand column below. Subsequent questions are then focussed on the more complex functions listed in the right column. The language functions listed are appropriate to the testing context, suitable for question writing and, in terms of linguistic difficulty, split in accordance with typical EFL teaching programs.

Simple (Concrete) Functions	**More Complex (Moving Towards Indefinite/Abstract) Functions**
Provide Information	*Express Opinion*
Describe	*Suggest/Advise*
Explain	*Speculate*
Compare	*Predict*

- Question sets should:
  - be written simultaneously to promote a fair balance of functions and approximate difficulty across each set
  - avoid inclusion of the same question (even differently-worded) if it exists in other sets within the handbook pairing.

- Each 7-question set should:
  - begin with the same question *Could you tell me about your job?*
  - include a question referencing the simple past
  - include a question referencing the future or a suggested future change to the role
  - employ as simple vocabulary as is possible referring to the *Word List*
  - employ as simple grammar as is possible
  - employ a variety of questions forms including:
    - open and closed questions*

        - a variety of question starting forms (What? How? Which? Can you? etc.)
    - *provide the interlocutor with a maximum of 2 extra-prompt options where questions may elicit a one-word (yes/no) or short response. These prompts are to be written in parentheses and are typically (*Why? / Why not?*), (*Could you tell me more?*) and (*In what way?*).

- The chart below shows how a balance of language functions are typically employed across 3 sets (1 handbook). In this case, the item writer's job is to write different questions (to complete the third column) based upon the proposed language functions:

Set A	Provide Information	*Could you tell me about your job?*
	Explain	
	Describe	
	Compare	
	Describe	
	Express Opinion	
	Suggest/Advise	
Set B	Provide Information	*Could you tell me about your job?*
	Describe	
	Explain	
	Compare	
	Provide Information	
	Express Opinion	
	Speculate	
Set C	Provide Information	*Could you tell me about your job?*
	Describe	
	Explain	
	Compare	
	Provide Information	
	Express Opinion	
	Predict	

Item Trialling & Item Revision

Once written sets are returned to the test development team (TDT), a process of in-house pre-production inspection takes place in order to decide if any obvious design faults can be eliminated through inspection. At this stage, it is also important to discard any items which are construct-irrelevant which, in this context, could be:

- questions which elicit inappropriate or non-assessable language (technical or operational language),
- questions which may unfairly ask candidates to respond to non-common, concrete and work-related topics they are unfamiliar with (e.g. a private pilot being asked about large international airports)

- questions which may appear too difficult – in terms of lexical choice or structure – for a 'warming-up task

The second stage of pre-testing analysis involves the delivery of question sets to groups of potential future test candidates (at test centres which are connected to training courses and, therefore, have willing candidates). The groups are selected to be as representative as possible of the future test-taking population in terms of role, gender, age, nationality, first language and benchmarked level. The minimum group size is 30. Sets are delivered (as per standard interlocutions guidelines) by TEA Examiners and the interviews are recorded for the TDT to later analyse with the following questions in mind:

- *Did the set elicit an assessable language sample?*
- *Was the set more difficult for this candidate than other sets?*
- *Was the set appropriately difficult for the candidate's level?*
- *Did the candidate have more problems answering the later questions than the earlier questions?*

and, of each question, the following 'tick-list' questions, with a *Comments* section with each:

- *repetition required?*
- *response?*
    - *appropriate?*
    - *approximate response time?*
    - *intended language function elicited?*
- *no response?*
    - *due to grammatical competence?*
    - *due to lexical competence?*
    - *due to sound / word / chunking recognition?*
    - *due to inappropriate theme for candidate?*

After the sets have been delivered by the interlocutor, the candidate is asked if he can explain the difficulties he had with certain questions.

Following the "response as evidence" edict described by Fulcher & Davidson (2010: 64), the data collated from these trials allows the TDT to assess the performance of each item and revise items where patterns of poor performance emerge. It was noted, for example, that the word 'routines' in the question *How important are routines in a pilot's job?* caused Russian speakers at lower levels difficulties as a result of mistranslation from the Russian language. Since a similar pattern of difficulty was not discovered among higher-level or non-Russian speakers, the item was neither dropped nor revised since, ultimately, the ability to distinguish between proficiency levels in plain English is the test's objective. Dropping or revising items is therefore, in the end, a judgement of the TDT based on the data, common sense, linguistic knowledge and experience, and not one based on intuition.

Part 2A

As stated above, the items for Part 2A are first written in line with the test-writing framework in the UK, before experienced operational pilots and controllers check the items for both operational authenticity and technical accuracy. However, the recording of items for all Part 2 tasks are produced in-house through the selection and employment of recording artists, meaning that a large bank of many thousands of items is available to the test development team.

To generate items for use in live testing, the stages of production are:

- item drafting + item checking 1
- item redrafting + item checking 2
- item recording + archiving

Item-writing Framework

Items for part 2A

- are monologues (from pilots and controllers)
- consist of 2 connected parts
- are between 10 and 20 words long
- are written in plain English yet are operationally accurate
- are set in non-routine aviation situations.

In writing a variety of different items, the following construct factors should be considered:

Domain:

D1	Health	(2)
D2	Technical	(2)
D3	ATC & Ground	(2)
D4	Weather & Environment	(2)
D5	Interference & Passenger Problems	(2)

Speaker:

S1	Pilot	(6/7)
S2	Controller	(2/3)
S3	Pilot or Controller (i.e. non-specific)	(0/1)

Function *(2parts to message, in any order)***:**

F1	Statement + Statement	(5)
F2	Statement + Request	(2)
F3	Statement + Question	(1/2)
F4	Statement + Command	(1/2)

Tense *(the most complex in either part of the message)***:**

T1	Present Simple	(3)
T2	Present Continuous	(1/2)
T3	Past Simple	(1/2)
T4	Past Continuous	(0/1)

T5	Future	(1)
T6	Present Perfect	(2)
T7	Conditional	(0/1)
T8	Passive	(0/1)

Modality *(any use of mood-modifying verbs)***:**

M1	Yes	(5/6)
M2	No	(4/5)

Negation *(in either part of message)***:**

N1	Positive elements only	(7)
N2	Negative elements	(3)

Contraction:

C1	Yes	(4/5)
C2	No	(5/6)

Certainty/Doubt *(in either part of message)***:**

D1	No Doubt	(5/6)
D2	Some Doubt	(4/5)

Phase of Flight:

P1	Ground	(4)
P2	En-route	(3/4)
P3	Non-specific	(3/4)

(The numbers in brackets detail the breakdown of factors appearing on each TEA CD – this is described in more detail below.)

Item writers are given lists of lexical sub-domains to prompt them, although they are not exhaustive. For example, for the domain of *Health*, the following list is provided and items are written to feature such vocabulary:

Ache	*Cuts and bruises*	*Medical*
Allergy	*Death*	*- assistance*
Ambulance	*Diabetes*	*Medical*
Asthma	*Dizziness*	*- emergency*
Birth	*Fainted*	*Medical update*
Bleeding	*First Aid*	*Nausea*
Breathing	*Food poisoning*	*Pain*
Broken bones	*Heart attack*	*Panic attack*
Burnt	*Heart problems*	*Passenger illness*
Calm	*Injury major*	*Pilot illness*
Consciousness	*Injury minor*	*Virus*

Following the initial writing stage, checking and re-drafting begins. Analysis of items by operational personnel occasionally gives rise to problems such as those exemplified below:

Item 1: *We had a fire in the left engine but put it out. We're now trying to get it started again.*

Expert view from Manuel, a Spanish pilot:
"This is unlikely since the pilot wouldn't know what caused the fire and wouldn't want to risk re-starting. In most cases, he would try to land."

Item 2: *You must abort take-off. There is smoke coming from your number 2 engine.*

Expert view from Paula, a Colombian controller:
"It is not the controller's job to tell the pilot what to do. We inform and then wait for his decision."

Item 3: *I'm getting technical help with my radar screen. I'll inform you when the situation is resolved.*

Expert view from Toni, a Swedish pilot:
"The only time I could think of something like this would be if you're flying to an airport with only one controller and he/she would be over informative that day. Normally they wouldn't bother you with that kind of information."

Item Production & CD Production

After item writing, the second stage of item production is recording. Voice artists are recruited to record all written items, helping to generate a large bank of items. Artists are selected for their accent (as influenced by their first language), the clarity of their pronunciation, and their language proficiency*. A mix of artists is employed in order to achieve the appropriate

balance of voices, considering gender and the predominant international accents. The two further construct factors generated through item recording are Voice and Accent:

Voice:

M	Male	(6)
F	Female	(4)

International Accent:

1	Native British	(1)
2	Native North American	(1)
3	Native Other	(0/1)
4	Spanish	(1)
5	French	(1)
6	Slavic	(1)
7	Arabic	(1)
8	Chinese	(1)
9	Italian	(0/1)
10	Portuguese	(0/1)
11	Germanic	(0/1)
12	Asian Other	(0/1)

*The concept of 'appropriate' international accents was a difficult one for the TDT to consider from the following points of view:

- *Authenticity*: training was provided in order to instruct the voice artists how to deliver the messages as naturally as possible with the appropriate tone and tempo.
- *Range*: By including a wide range of international accents on every test CD, attempts were made to not unfairly benefit or penalise test candidates of particular demographics who had greater exposure to a narrower range of accents. In assessment terms, allowances are made for lapses in comprehension: even at Level 6, 100% comprehension is not essential. (See ***Report 11 - Examiner Training & Assessment using TEA*** for more information.)
- *Clarity*: The ICAO Descriptors state under Comprehension Level 4 that "comprehension is often accurate on common, concrete, and work related topics when the accent or variety used is sufficiently intelligible for an international community of users". It was agreed that the voice artists should have a high level of pronunciation proficiency themselves –Level 5 or Level 6 – to be 'sufficiently intelligible' since:
    - in language testing terms, it would be unfair to assess listening comprehension against recordings of 'inaccurate' pronunciation
    - in the international aviation context, pilots and controllers need to understand native and near-native speakers of English (the highest proficiency levels).

After recording, TEA CDs are constructed. The first step is to select messages that can co-exist in one Part 2A based on a template of the construct factors. The numbers in brackets above indicate the frequency of appearance in each Part 2A set. So in terms of *Phase of Flight*, for example, each 2A set includes 4 messages based on ground and airport operations,

3 or 4 messages set in en-route situations, and 3 or 4 messages in which the context is non-specific.

One can see that each Part 2A generates a balance of linguistic features through a set template in order to balance the potential linguistic challenges of each test CD. Furthermore, attempts to balance the sets in this way ensures that the construct factors intended to engage the language competences to be measured by TEA (see ***Report 01 – Description of Test Purpose, Specifications and Construction*** for further information) are forced into every test since aspects of tense, modality, contraction, etc. are equivalent on each CD.

Statements were written in order to rationalise the balance of factors into each CD template:

- a variety of situations from the general topic domains are covered equally
- a variety of situations to reflect different phases of flight are equally covered
- non-routine messages are more likely to be delivered by pilots than controllers
- male voices are more prevalent than female voices in aviation
- situational statements are more common than requests, questions or commands
- positive statements are more common than negative statements
- non-routine situations are more likely to describe a present situation than a past/future one
- more proficient speakers use contractions in plain English than less proficient speakers

Domain	Sub-domain	Message	Difficulty	Pilot / ATC	Function	Tense	Modality	Pos/neg	Contraction	Doubt	Phase
Weather	Visibility Better	*The aerodrome is now in sight. We can continue visually.*	1	P	St-Statement	Pres Simp	Yes	Positive	No	No	Ground
Technical	Fire & Smoke	*We're evacuating because of smoke. Call the fire brigade.*	1	P	St-Command	Pres Cont	No	Positive	Yes	No	Ground
Interference	Radio Problem	*I couldn't catch what you said due to interference. Could you repeat the last part of your message?*	1	P/C	St-Request	Past Simp	Yes	Negative	Yes	Some	Non-Specific
Health	First Aid	*Our take-off is going to be delayed because a passenger is receiving some first aid.*	2	P	St-Statement	Future	Yes	Positive	No	Some	Ground
ATC/Ground	Airport Problems	*Your destination airport is closed because they're having runway clearance problems.*	1	C	St-Statement	Pres Cont	No	Positive	Yes	No	Non-Specific
Interference	Traffic	*The traffic which just crossed our heading left to right was too close. What's happening?*	2	P	St-Question	Past Simp	No	Positive	No	Some	En-route
Weather	Thunderstorm	*There's a thunderstorm ahead. You need to turn right to an easterly heading.*	1	C	St-Statement	Pres Simp	Yes	Negative	Yes	No	En-route
Technical	Engine	*We've tried to restart the engine but it's not responding.*	2	P	St-Statement	Pres Perf	No	Negative	Yes	No	Non-Specific
Health	Heart problems	*We've got a passenger on board with heart problems. Request emergency descent for priority landing.*	1	P	St-Request	Pres Perf	No	Positive	Yes	No	Ground
ATC/Ground	Navigation	*There is an airport with a tower eight miles south of your position. Do you have enough fuel?*	3	C	St-Question	Pres Simp	Yes	Positive	No	Some	En-route

The image on the left displays the 2A set that was developed for use in TEA CD17.

Each message is also given an assumed difficulty rating (the red column here) from 1 (easier) to 3 (more difficult) by the TDT in order to account for assumed difficulty prior to trialling with sample candidates. In line with best test practice, recordings that are found to be easier are put at the beginning of each CD in order to ease the candidate into the task as much as possible. In each case, the message assumed to be most difficult – although trialling sometimes disproves it – appears as the final item in 2A.

Following the completion of each 2A set on paper and in line with the CD construction template, 'voices' (male-female, accented) are randomly assigned to each message and recordings from the item bank are gathered to produce a 'trial CD'. The image below shows the breakdown of voices across TEA CD17, including the items for Parts 2B (in blue) and 2C (in green). It can be seen that candidates are exposed to a wide range of international accents during TEA.

1	The aerodrome is now in sight. We can continue visually.	Male	Spanish
2	We are evacuating because of smoke. Call the fire brigade.	Male	Portuguese
3	We had radio interference and couldn't hear you. Could you repeat your last message?	Female	Native British
4	Our take-off is going to be delayed because a passenger is receiving some first aid.	Male	Germanic
5	Your destination airport is closed because they're having runway clearance problems.	Female	Italian
6	The traffic which just crossed our heading left to right was too close. What's happening?	Male	Arabic
7	There is a thunderstorm ahead. You need to turn right to an easterly heading.	Female	French
8	We've tried to restart the engine but it's not responding.	Male	Native Other
9	We have got a passenger on board with heart problems. Request emergency descent for priority landing.	Female	Chinese
10	There's an airport with a tower eight miles south of your position. Do you have enough fuel?	Male	Asian Other 2
11	We have a problem.... Some of the passengers are missing.	Female	Slavic
12	We have a situation... The passenger's not conscious.	Female	Native North American
13	We have a problem.... An alarm has just started.	Male	Italian
14	We need some help.... We have a woman here who is very upset.	Female	Arabic
15	We need some help.... We can't get the computer to work.	Male	Native Other
16	We need some help.... A man's bags have been stolen.	Female	Germanic

Item Trialling & Item Revision

Test CDs are produced in batches of 4 (for each Examiner Handbook) and the trialling is conducted in the same manner to achieve equivalence as far as possible. See ***Report 09 – Assessing the Reliability of Test CDs during Development*** for more information about CD equivalence.

The trialling and revision process is as follows:

- Native speaker trials for speaker intelligibility: a small group of native speakers check the recordings, transcribing the messages word for word; where patterns of requiring repetition or item difficulty emerge, recordings are immediately changed (the chosen 'voice' is swapped)
- Then, 2 groups of 10 candidates (different nationalities, aviation positions, gender, levels) are generated
- each candidate is given a Control CD (to award a TEA Comprehension level)
- candidates are tested with 4 CDs across 2 sessions (to avoid weariness) using the counter-balanced delivery approach
- after each trial, items which proved difficult are listened to & discussed
- recordings are listened to by Senior TEA Examiners, scores are awarded for performance, and results are analysed
- items and item sets are revised where:

- items are not discriminating well
- items are too easy or too difficult
- items are positioned poorly on a CD

▪ revisions may include:
- 'swapping' voices if the trialled voice appears to influence the item negatively
- slowing the playback speed
- dropping an item in preference of another that matches the CD template
- re-ordering items within the CD template

- the revised CD is then given to a third group of 10 different candidates and the process is repeated:
  - Item performance is monitored
  - facility values and discrimination indices are calculated to assess the need for further modification before operalisation (in live testing)

Through the analysis of facility values and discrimination indices, amendments of both items and CDs are made to try to balance both the levels of difficulty and differentiation while retaining the set criteria for balanced CD construction.

For example, the data below is taken from initial trialling of CD1 with 2 groups of sample candidates. Correct responses were scored 1, incorrect 0. The facility value (F.V.) shows how difficult an item was for the group as a whole:

Group	Candidate	TEA Comp	Nationality	1	2	3	4	5	6	7	8	9	10	Score
A	1	6	Bulgarian	1	1	1	1	1	1	1	1	1	1	10
	2	5	French	1	0	1	1	1	1	0	1	1	0	7
	3	5	Colombian	1	1	1	1	1	1	1	1	0	0	8
	4	5	Bulgarian	1	0	1	1	1	1	1	1	1	0	8
	5	4	Italian	1	1	1	1	1	0	1	1	1	0	8
	6	4	Italian	1	1	1	1	1	0	1	1	1	0	8
	7	4	Polish	1	0	0	1	1	1	1	0	1	0	6
	8	3	Russian	1	0	0	0	0	0	0	0	0	0	1
	9	3	Brazilian	1	0	0	0	1	0	0	0	0	0	2
	10	3	Russian	1	0	0	0	0	0	0	0	0	0	1
B	11	6	Bulgarian	1	0	1	1	1	1	1	1	1	0	8
	12	5	Bulgarian	1	0	1	1	0	1	1	1	1	0	7
	13	5	Spanish	1	0	1	1	1	1	1	1	0	0	7
	14	5	Bulgarian	1	0	1	1	1	1	1	1	1	0	8
	15	4	French	1	0	1	1	1	1	0	1	1	0	7
	16	4	Spanish	1	0	1	1	0	1	0	1	1	0	6
	17	4	Italian	1	0	1	0	0	0	1	1	0	0	4
	18	3	Polish	1	1	1	0	0	0	0	0	0	0	3
	19	3	Russian	1	0	1	0	0	0	1	0	0	0	3
	20	2	Russian	1	0	1	0	1	0	0	0	0	0	3
Facility Value Per Item (100=easy)				100	25	80	65	65	55	65	65	55	5	58%

From the image, left, item 1 has an F.V. of 100 meaning that every candidate got the item correct: it does not differentiate between ability at all.

Item 10 has an F.V. of just 5 as only 1 candidate (a Level 6 candidate) gave a correct response. It might be too difficult but does differentiate at the highest level.

	Candidate	TEA Comp	Nationality	1	2	3	4	5	6	7	8	9	10	Score
	1	6	Bulgarian	1	1	1	1	1	1	1	1	1	1	10
	3	5	Colombian	1	1	1	1	1	1	1	1	0	0	8
	4	5	Bulgarian	1	0	1	1	1	1	1	1	1	0	8
	5	4	Italian	1	1	1	1	1	0	1	1	1	0	8
	6	4	Italian	1	1	1	1	1	0	1	1	1	0	8
	11	6	Bulgarian	1	0	1	1	1	1	1	1	1	0	8
	14	5	Bulgarian	1	0	1	1	1	1	1	1	1	0	8
	2	5	French	1	0	1	1	1	1	0	1	1	0	7
	12	5	Bulgarian	1	0	1	1	0	1	1	1	1	0	7
	13	5	Spanish	1	0	1	1	1	1	1	1	0	0	7
	15	4	French	1	0	1	1	1	1	0	1	1	0	7
	7	4	Polish	1	0	0	1	1	1	1	0	1	0	6
	16	4	Spanish	1	0	1	1	0	1	0	1	1	0	6
	17	4	Italian	1	0	1	0	0	0	1	1	0	0	4
	18	3	Polish	1	1	1	0	0	0	0	0	0	0	3
	19	3	Russian	1	0	1	0	0	0	1	0	0	0	3
	20	2	Russian	1	0	1	0	1	0	0	0	0	0	3
	9	3	Brazilian	1	0	0	0	1	0	0	0	0	0	2
	8	3	Russian	1	0	0	0	0	0	0	0	0	0	1
	10	3	Russian	1	0	0	0	0	0	0	0	0	0	1
Facility Value Per Item (100=easy)				100	25	80	65	65	55	65	65	55	5	58%

In the image, left, the candidates have been ranked by performance. Items with F.V. values between 20 and 80 seem to be differentiating well.

Discrimination index (D.I.) shows how well a question differentiates between high and low scorers. You would expect that high scoring students would select the correct answer for each question more often than low scoring students. If this is true, then the assessment is said to have a *positive discrimination index* (between 0 and 1) - indicating that candidates who received a high total score chose the correct answer for a specific item more often than the candidates who had a lower overall score. If, however, you find that more of the low-performing candidates got a specific item correct, then the item has a *negative discrimination index* (between -1 and 0).

From the patterns of 1s and 0s in the image above, it can be seen at-a-glance that items were generally answered correctly as expected i.e. the higher-level candidates understood the items at the beginning of the task (see the abundance of 1s grouped in the top left-hand corner) except for problems with item2; and the lower-level candidates generally misunderstood the items at the end of the task (see the 0s in the bottom right-hand corner).

	F.V.	D.I.
Item1	100%	0.0
Item2	25%	0.5
Item3	80%	0.5
Item4	65%	1.0
Item5	65%	0.7
Item6	55%	0.7
Item7	65%	0.8
Item8	65%	1.0
Item9	55%	0.8
Item10	5%	0.2

To calculate the D.I., the top and bottom thirds of the ranked candidates are separated (those highlighted in purple in the image above) into groups. The number of students in the lower group who got the item correct is subtracted from the number of students in the upper group who got the item correct. The result is then divided by the number of students in each group.

The F.V.s and D.I.s for this trial with CD1 are recorded in the table on the left. We can see that most of the items are neither too easy nor too difficult and are discriminating well (high, positive D.I. values).

Questions for further investigation (leading to revision before trialling with a third group of candidates) that arose for the TDT from this data were:

- is there a problem with either the production or the assessment of item2?

- what is it about item6 that prompted all 3 Italian candidates to misunderstand it?
- is item10 too difficult for the average very high-level candidate?
- is item1 too easy for the average very low-level candidate?

In a test that aims to assess a wide-range of proficiency levels, difficult items and items which differentiate widely are important features to consider in pre-testing. All TEA CDs are subjected to analysis of this nature.

Parts 2B & 2C

Items for Parts 2B & 2C are written in line with the test-writing framework and, as with Part 2A, are then recorded in-house through the selection and employment of recording artists, meaning that a large bank of many thousands of items is available to the test development team.

Item-writing Framework

Items for parts 2B & 2C

- are monologues (from aviation personnel in general)
- are written in plain English
- always begin with a stock introductory phrase to allow the candidate to 'tune in' to the voice (for 2B, either "*We have a problem…*" or *"We have a situation…"* to elicit questions; for 2C, *"We need some help…"* to elicit advice)
- are short messages describing a problem in unexpected/unusual aviation situations (but not necessarily in the cockpit/control tower).
- in 2B, the messages are de-contextualised to generate the information gap necessary to encourage the asking of questions to find out further information
- in 2C, the messages are more specific to generate the need for appropriate suggestions and advice from candidates.
- item writers should ask themselves whether the candidate would be likely to be able to respond sensibly in his own first language. In that way, designing items which elicit suitable language samples while conforming to the focus of a 'broad' work-related context is best managed.

In writing a variety of different items, the following construct factors should be considered:

Domain:

D1	Environmental	(1/2)
D2	Health	(1/2)
D3	Human	(1/2)
D4	Technical	(1/2)

Tense :

T1	Present Simple	(3)
T2	Present Continuous	(1/2)
T3	Past Simple	(0/1)
T4	Past Continuous	(0/1)
T5	Future	(0/1)
T6	Present Perfect	(1)
T7	Conditional	(0/1)
T8	Passive	(0/1)

Modality :

M1	Yes	(1/2)
M2	No	(4/5)

Negation:

N1	Positive elements only	(4/5)
N2	Negative elements	(1/2)

Contraction:

O1	Yes	(3)
O2	No	(3)

Prior to recording, items are reviewed by the TDT and trialled with a sample group to informally assess the potential for elicitation of each item. Those considered 'weak' items are dropped from the recording process. The recording artists described in the Part 2A section are then trained to record and archive the items.

To ensure a balance of items across the tasks, the CD production template for Parts 2B & 2C decrees which items can co-exist on one test CD – the numbers in brackets above indicate the frequency of appearance in each Part 2B/C set (of 6 recordings). Voices are also balanced:

Voice:

M	Male	(3)
F	Female	(3)

The content of Part 2A items is also considered when selecting items for a CD as the TDT does not want repetition of vocabulary or situations on the test CD. And the 'accents' are selected to balance those used in Part 2A and avoid repetition as far as possible.

Item Trialling & Item Revision

Trial CDs are trialled in 2 ways:

- native speaker trials are conducted to check the intelligibility of the items
- trials are conducted with the sample groups as described in the Part 2A section above.

From the resulting data collected during the trials, the 2B and 2C sets are analysed as one task and those which are unable to differentiate appropriately between candidates of higher level of comprehension are revised (candidates at Level 4 and 5 should have some problems understanding a couple of recordings since the items are set in unexpected situations or contain linguistic difficulties). Items which are too difficult for the highest-level candidates are re-produced using different messages and/or voices.

Part 3 Pictures

The part 3 tasks (*Pictures* and *Discussion*) are linked by theme and the themes run across the pairs of Examiner Handbooks in order to vary and balance test content from one handbook to the next. There are 12 themes, 6 per handbook:

- *Aerodromes*
- *Airports & the Environment*
- *Aviation Growth*
- *Dangers*
- *Emergencies & Safety*
- *Health*
- *Organisations & Training*
- *People & Communication*
- *Security*
- *Technology*
- *Time & Schedules*
- *Weather & Geography*

Item-writing Framework

Item 'writers' search for photographs of aviation situations which can complement each other allowing in the following ways:

- a link to the theme (e.g. security checks at airports under the theme *Security*)
- a clear link between the pictures (e.g. passengers being searched manually and passengers' bags being sniffed by police dogs)

Pictures which depict unusual situations are superior since a wider range of language is elicited, and candidates are more likely to speculate and suggest if it is not clear what has happened. For that reason, images of infamous aviation scenes are best avoided, as are 'disaster' scenes.

Pairs of pictures should then meet the following criteria:

- clarity of image
- contains multiple information – a range of detail that can be described (either by known lexis or paraphrase) but cannot be described fully in one or two sentences
- depict actions in progress in order to elicit a range of grammatical structures
- depict both similarities and differences (e.g. manual security searches versus electronic security checks) to potentially elicit a range of comparative language.

Official images, names and logos are removed from chosen images.

An example set:

This picture set works well because:

- they are clearly linked aviation situations (*emergency evacuation*)
- each picture depicts activity, people and objects
- there are clear similarities and differences between them
- they may elicit a range of vocabulary (*rope, slide, lean out, go down*)
- they elicit speculation (*training or real situation?*; *why is one of the flight crew leaving the plane from the window?).*

Item Trialling & Item Revision

Trials are conducted with a range of sample candidates. For each picture set (of 2 pictures), a minimum of 20 candidates have their descriptions recorded for later analysis by the TDT with the following questions in mind:

- *Did the set elicit an assessable language sample?*
- *Did the candidate talk for one minute?*
- *Did the candidate talk for a period of time equivalent to his proficiency level?*
- *Were a range of functions elicited?*
- *Were a range of structures elicited?*
- *Did the candidate have problems finding the vocabulary he needed?*
- *Did the candidate need to paraphrase?*
- *Did the candidate have obvious problems linking the pictures?*

- *Did either picture cause the candidate non-linguistic difficulties?*

Where individual pictures or sets do not generate the intended response, items are changed or dropped. The following pictures were dropped for the reasons given:

Responses to this picture demonstrated that there was not enough detail or activity depicted to elicit a full language sample. Furthermore, it was not clear whether a lack of cultural knowledge was causing candidates to stall unfairly.

Candidates' responses to this picture indicated that there was not enough clear activity visible. There was little to describe after mentioning 'a busy security scene at an American terminal'.

Part 3 Discussion

Item-writing Framework

Since the discussion is intended to elicit language in 'broader' aviation contexts (see ***Report 03 - Description of Tasks & Instructions*** for more details), items may move into more structurally and lexically complex areas than the rigid work-related format of Part 1. Items are written in sets of 3 questions under sub-themes of the 12 themes listed above. For example, under the theme *Time and Schedules*, the following sub-themes may be considered:

- *Delays/Affect of delays/Solutions to delays*
- *Passenger/Baggage problems*
- *Control of traffic*
- *Route planning*
- *Use of time for navigation / separation*
- *Last-minute delays*
- *Economics of flight and fuel*
- *Company pressures on operations*

- *Cargo versus Passenger flights*
- *Time pressures*
- *Baggage handling / Cargo organisation*

Each theme contains 9 scripted questions – 3 sub-sets of 3 questions in each. Items are written with the targeted language functions (below) in mind to be varied and balanced across the 9 questions:

Question 1	Question 2	Question 3
Provide Information	*Express Opinion*	*Speculate*
Describe	*Elaborate (on a prompt)*	*Predict*
Explain	*Compare*	*Suggest/Advise*
	Reassure	*Suggest*

As with Part 1 item-writing, initial questions focus on simple, concrete functions moving to functions of a more complex, abstract nature as the set progresses.

- Each question set should employ a variety of questions forms. Allowance for open and closed questions is free since interlocutors are trained and expected to extend candidates' responses in this task. Although items should be written in as simple language as possible, difficulties with lexis and structures are acceptable here since candidates at higher levels are expected to be able to manage the interaction and negotiate understanding with the interlocutor as necessary. Item writers should ask themselves whether the candidate would be likely to be able to respond to the question easily, sensibly and fully in his own first language. In that way, designing items which elicit suitable language samples while conforming to the focus of a 'broad' work-related context is best managed. Candidates at lower-levels are not expected to be able to discuss these broader aviation topics at length without the help of the simpler first question and interlocutor prompting.

Here is an example set of 9 questions under the theme of *Weather & Geography*:

Handbook	Set	Sub Topic	Function	Question
			Describe	*Which weather conditions are bad for pilots and controllers?*
	1	Bad Weather Conditions	Reassure	*Passengers are sometimes worried about bad weather. Should they be worried?*
			Speculate	*How could aviation companies make passengers more relaxed about bad weather conditions?*
			Explain	*Which climates are best for aviation?*
E	2	Climates	Express Opinion	*Do you think the aviation industry has a negative effect on our environment?*
			Predict	*How will attitudes to climate change affect the aviation industry in the future?*
			Provide Information	*Where does information about weather conditions in aviation come from?*
	3	Weather Information	Compare	*What is the difference between how controllers and pilots receive news about weather?*
			Suggest	*At what point are flights diverted because of expected weather problems?*

One can see that there is a variety of functions and question forms across the set.

Item Trialling & Item Revision

Once written sets are returned to the test development team (TDT), a process of in-house pre-production inspection takes place in order to decide if any obvious design faults can be eliminated through inspection. At this stage, it is also important to discard any items which are construct-irrelevant which, in this context, could be:

- questions which elicit inappropriate or non-assessable language (technical or operational language),
- questions which may unfairly ask candidates to respond to narrow, specific aviation themes which requires knowledge they cannot even speculate about (e.g. a controller being asked about flying a plane)
- questions which may be written more simply, in terms of lexical choice or structure

The second stage of pre-testing analysis involves the delivery of question sets to groups of potential future test candidates (at test centres which are connected to training courses and, therefore, have willing candidates). The groups are selected to be as representative as possible of the future test-taking population in terms of role, gender, age, nationality, first language and benchmarked level. The minimum group size is 30. Sets are delivered (as per standard interlocutions guidelines) by TEA Examiners and the interviews are recorded for the TDT to later analyse with the following questions in mind:

- *Did the set elicit an assessable language sample?*
- *Was it possible to develop a natural discussion from the scripted questions?*
- *Did the candidate have more problems answering the later questions than the earlier questions?*

and, of each question, the following 'tick-list' questions, with a *Comments* section with each:

- *repetition required?*

- *response?*
    - *appropriate?*
    - *approximate response time?*
    - *intended language function elicited?*

- *no response?*
    - *due to grammatical competence?*
    - *due to lexical competence?*
    - *due to sound / word / chunking recognition?*
    - *due to inappropriate theme for candidate?*

After the sets have been delivered by the interlocutor, the candidate is asked if he can explain the difficulties he had with certain questions.

Following the “response as evidence” edict described by Fulcher & Davidson (2010: 64), the data collated from these trials allows the TDT to assess the performance of each item and revise items where patterns of poor performance emerge. Final decisions on dropping or revising items are, in the end, a judgement of the TDT based on the data, common sense, linguistic knowledge and experience, and not one based on intuition.

D – Version Content

As this report has described, every effort is made to check item quality and structure items within tasks in order to make different TEA versions (sets within tasks, tasks within handbooks) as equivalent as possible. While recognising that the production of multiple, truly-parallel versions is an unattainable concept in language testing, the frameworks for handbook, task and item development help to make version content across TEA forms as fair and comparable as possible.

Set frameworks: Allow for the reproduction of like-for-like forms

Thematic balance: Themes and sub-themes are included a maximum of once in any Examiner Handbook. This is also true for items for Part 2 where multiple references to the same type of non-routine situation are avoided.

Functional balance: Within question sets, a balance of items intended to elicit a variety of language functions is considered.

Lexical balance: Within the test CDs, lexical domains are covered equally and vocabulary is not repeated.

Equivalent levels: Every effort is made to ensure that question sets & test CDs do not unfairly hinder or help a candidate's performance and jeopardise test reliability. See ***Report 09 – Assessing the Reliability of test CDs during Development*** for further information.

Reference

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