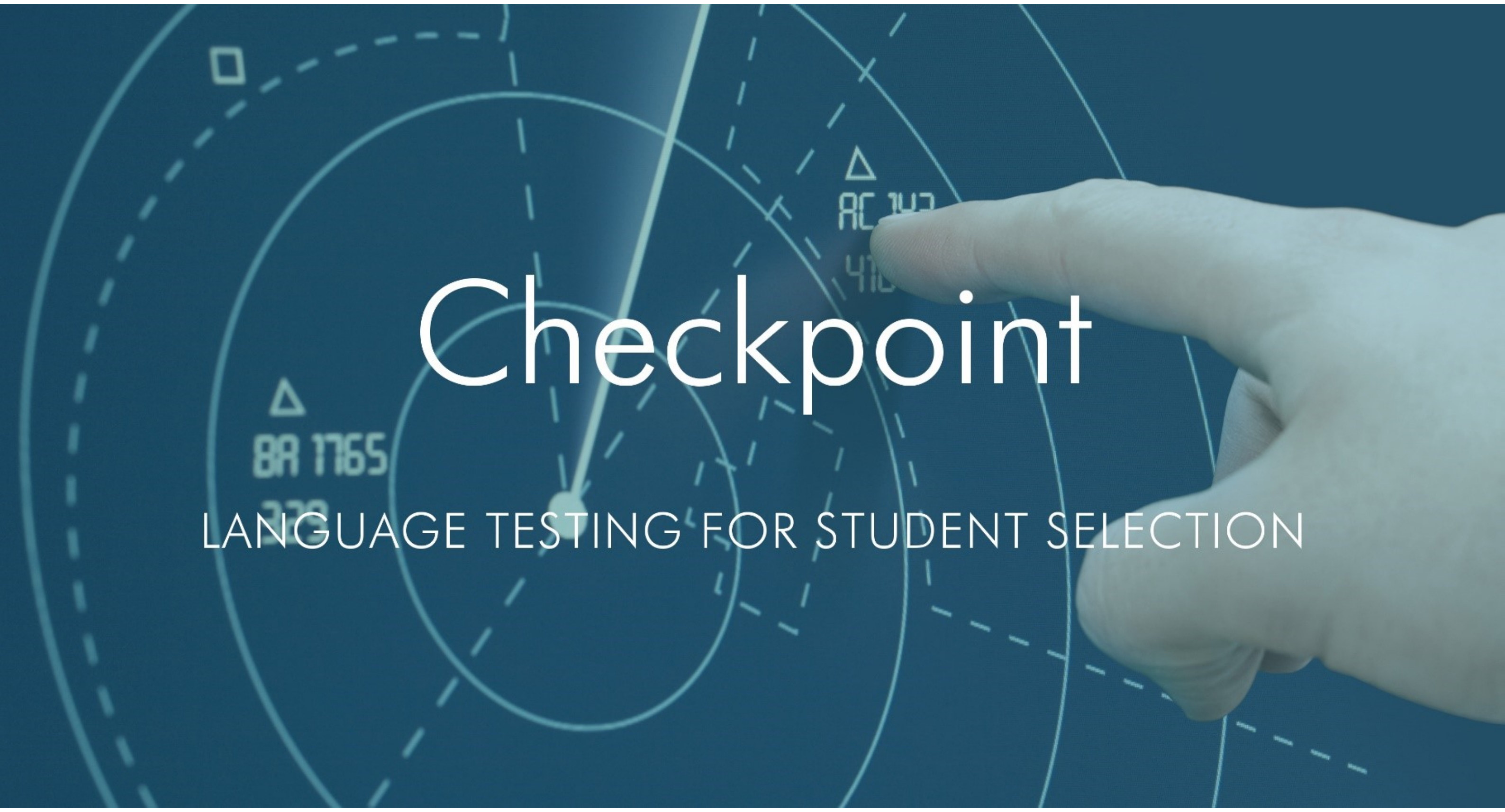
Latitude
Aviation English Services



Checkpoint

LANGUAGE TESTING FOR STUDENT SELECTION

Test structure, platform and scores

May 2017

Contents

1. Introduction

Checkpoint is an English Language Proficiency (ELP) test designed to measure the English language skills required for successful English-medium aviation training. Checkpoint is owned and operated by Latitude Aviation English Services Limited (UK).

At Latitude, we are committed to providing quality language training and testing products and services, and to helping our stakeholders to understand and use them. This document provides information on the test structure, the administration platform and test scores. The document is intended primarily to help aviation training decision-makers and admissions officers decide if Checkpoint meets their language testing requirements, but it may be of interest to other stakeholders in aviation training such students, student sponsors, English language instructors, aviation assessors and training managers.

2. Test description

Checkpoint is a specific-purpose web-based test of ELP designed to help airlines, Air Navigation Service Providers (ANSPs) and civil and military Aviation Training Organisations (ATOs) select students for ab-initio flight and Air Traffic Control (ATC) training and make decisions about admission to aviation training programmes.

Checkpoint is designed to be used:

- Before student assessment conducted in the medium of English;
- Before or after student assessment conducted in the mother tongue; and/or
- At the end of an English language training course.

Checkpoint test scores are designed to align to the Common European Framework of Reference (CEFR) and can be used alongside existing skills assessment procedures to:

- Determine student language proficiency for entry to ab-initio pilot / ATC training programmes; and
- Identify any student language training requirements.

3. Test platform

Checkpoint is administered via a specific-purpose computer based language testing platform developed and operated by Owl Testing Software, Pittsburgh, USA (www.owlts.com).

The Owl Test Management System (TMS) is an extremely versatile and flexible flash-driven platform that allows for the administration of large scale high stakes computer based language testing programmes. The Owl TMS centralises test content and data and manages user access to the system according to pre-defined roles allowing:

- Latitude to create, administer and monitor tests, and assess candidate performance
- Latitude's customers to use Checkpoint from any location with a stable internet connection

Owl's clients include:

- The National Board of Certification for Medical Interpreters
- The College of Staten Island / City University of New York
- The Information and Communications Technology Council
- Colombian Ministry of Education
- Yale, Cornell and Columbia Universities

With thousands of test takers worldwide each year, the Owl TMS has a proven track record in reliable and robust language test delivery and management.

4. Test structure

NOTE: Detailed test description and task familiarisation videos for candidates are available on the Latitude website.

Part 1: Listening (Total time: 40 minutes including test introduction, test and task instructions and example items)

Task	Discourse type	Task time (minutes)	Speakers	Text length		Number and type of scored items
				Words	Minutes	
1	Informal student–training centre staff dialogue	11’	2+	1000 (+/- 100)	5-6	8 x 4-option MCQ (Answer the question / Complete the sentence)
2	Informal student-student dialogue	11’	3+	1000 (+/- 100)	5-6	As above
3	Formal training: instructor monologue with some instructor-student interaction	11’	1+	1000 (+/- 100)	5-6	As above

Part 2: Reading (Total time: 40 minutes including test and task instructions and example items)

Task	Discourse type	Task time (minutes)	Text length	Number and type of scored items
1	Extract from ICAO Doc 4444: Procedures for Air Navigation Services	7	300 (+/- 50)	4 x MCQ (True, false or not given)
2	Extract from FAA Aeronautical Information Manual	7	300 (+/- 50)	4 x 4-option MCQ (Complete the sentence)
3	Extract from UK Air Accident Investigation Branch incident report	12	600 (+/- 100)	8 x 4-option MCQ (Answer the question / Complete the sentence)
4	Extract from industry journal on aviation training, safety and management	12	600 (+/- 100)	8 x single-option MCQ (Paragraph matching)

Part 3: Speaking (Total time: 10 minutes including test and task instructions)

Task	Title	Task time (minutes)	Task description	Response preparation?	Response (seconds)
1	Animation description	4	The candidate describes a one-minute animation of an aeronautical mechanism or process	Yes. The candidate watches the animation twice before describing	80
2	Storyboard narration	2.5	The candidate describes an illustrated storyboard of an incident/accident in aviation	Yes. The candidate has one minute to look at the storyboard before narrating	80
3	Interview	2.5	The candidate reads/listens to and answers 3 questions on their future career in aviation and the aviation industry in general	No. The candidate responds to the questions as they are presented	120

5. Test scores

5.1 ICAO and the CEFR

Unlike most aviation language tests, Checkpoint does not measure language proficiency according to the ICAO Rating Scale. This is because the ICAO Rating Scale is an inappropriate measure of language proficiency for ab-initio aviation students for two key reasons:

1. *The ICAO Rating Scale addresses only spoken language (speaking and listening); it does not address reading … skills*[1]*.* As reading is a skill crucial to successful ab-initio aviation training, measurement of student language proficiency according to the ICAO Rating Scale is necessarily under-representative of the language skills required for successful ab-initio aviation training.

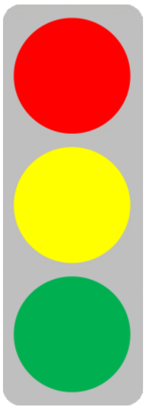
2. *The sole object of ICAO language proficiency requirements is aeronautical radiotelephony communications*[2]. Students on entry to professional aviation training are very unlikely to possess working knowledge of flight operations or experience with standard radiotelephony (RT) communications. Therefore, measurement of student language proficiency using tests designed to meet the ICAO language proficiency requirements is a threat to both test fairness and the validity and reliability of language test scores.

In seeking a more valid scale of measurement for Checkpoint, Latitude conducted research[3] into student ELP requirements and the suitability of the Common European Framework of Reference (CEFR) for language assessment in the context of ab-initio aviation training. In summary, this research involved linking aviation instructor's perceptions of the minimum levels of ELP required by students to CEFR reading and listening tasks and associated descriptors. The results showed that:

- The CEFR contains descriptors of ELP that are relevant to the context of ab-initio aviation training; and
- CEFR B2 describes a minimum entry-level of ELP for English-medium aviation training.

5.2 The Checkpoint traffic light system

Checkpoint scores are reported using a traffic-light system as follows:



Red: Language is likely to be an obstacle to successful aviation training for candidates that score red in any part of the test. We recommend that candidates who score red in any part of the test undergo 200+ hours of language training before beginning aviation training.

Yellow: Candidates that score yellow in any part of the test may encounter language-related difficulties during aviation training. We recommend that candidates who score yellow in any part of the test undergo 25-200 hours of language training before beginning aviation training.

Green: Candidates that score green in all parts of the test are unlikely to encounter language-related difficulties during aviation training.

[1] ICAO document 9835, *Manual on the implementation of Language Proficiency Requirements*, Section 4.5.5.a

[2] Ibid, section 3.2.7

[3] A full discussion of the issues associated with using the ICAO Rating Scale for measuring student ELP, along with presentation of the research summarised above, can be found in Emery, H. (2016) *Aviation English For The Next Generation* in Enright, A. and Borowska, A. (eds.) (2016) *Changing Perspectives on Aviation English Training*, Studi Naukowe 29, University of Warsaw.

5.3 The traffic light system and the CEFR

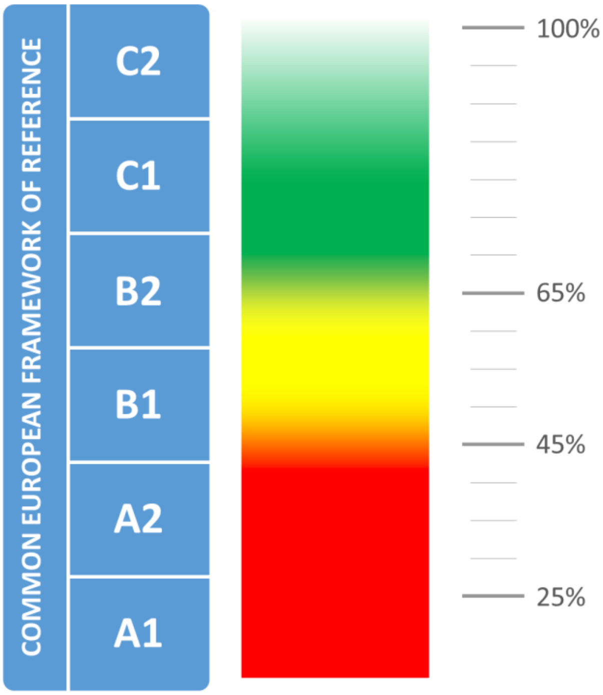
Checkpoint point scores are designed to align with the Common European Framework of Reference. This means that, in each of the Checkpoint listening, reading and speaking tests:

- Candidates who score red may be able to perform the tasks described at the A1 and A2 levels, but have a poor chance of performing the easier tasks described at the B1 level.

- Candidates who score yellow have a good chance of performing the easier tasks described at B1 and, depending on scores, may have a good chance of performing the harder tasks at B1 and a reasonable chance of performing the easier tasks at B2.

- Candidates who score green have a good chance of performing the easier language tasks described at B2 and, depending on scores, may have a reasonable to good chance of performing language tasks at C1 and C2.

The diagram below shows how Checkpoint scores are designed to align to the CEFR. The percentages link to listening and reading test scores where 45% and 65% represent cut-scores for the red-yellow and yellow-green levels respectively.



5.4 Listening, reading and speaking scores

Checkpoint listening and reading tests are scored automatically. Candidates receive a traffic light score and a percentage score for each of the listening and reading tests. During the speaking test, the candidate's voice is recorded by the computer for subsequent rating by Latitude's raters. Spoken performance is rated according to the Checkpoint rating scale for speaking. The rating scale is comprised of three levels – red, yellow and green (arranged vertically) and five criteria – task fulfilment, pronunciation, structure, vocabulary and fluency (arranged horizontally). Raters award each candidate a level in each criterion. The candidate's overall speaking score is the lowest of any score in the five criteria.

5.4.1 Task fulfilment

Task fulfilment focuses on how well the candidate addresses the requirements of the speaking tasks. In task one, the candidate describes how an aeronautical process or mechanism works based on an animation or a sequence of pictures. Rating task fulfilment in task 1 involves judging:

- The quality and accuracy of the candidate's description and how much of the visual and numerical information the candidate includes in their response; and
- How well the candidate incorporates the technical lexis presented in the animation / pictures in their response, and how accurately such technical lexis is used.

In task 2, the candidate provides a narrative based on a sequence of pictures. Rating task fulfilment in task 2 involves judging:

- The extent to which the candidate formulates a coherent narrative from the sequence of pictures; and
- The quality and accuracy of the descriptions of the visual information in the prompt.

Task fulfilment in task 3 relates to the degree to which the candidate's responses to the questions address the topics raised in the questions, and the level of detail, reasons and examples that the candidate provides as evidence to support their ideas.

As the three speaking tasks are designed to elicit a specific performance, any language which does not address the task requirements is considered irrelevant.

5.4.2 Pronunciation

Pronunciation focuses on how well the candidate can produce the features of the English sound system and the extent to which control of these features assists or impedes raters' understanding of the candidate. These features include:

- Production of individual vowel, diphthong and consonant sounds;
- Pronunciation of words with the correct syllable stress;
- Rise and fall of voice pitch (intonation) to show meaning, for example, certainty, emphasis, query, digression, conclusion etc; and
- Control of word stress, cadence and pausing to organise speech into meaningful chunks and to indicate the beginning, middle and end of units of speech.

Note: Candidates are not penalised for mispronunciation of the technical lexis presented to the candidate in task 1 as they may be encountering these words for the first time.

5.4.3 Structure

Structure focuses on the range of grammar the candidate uses and how accurate the candidate's grammar is. Rating structure involves identifying:

- Basic[4] and complex[5] structures;
- The extent of the range of grammatical structures used, i.e. how much flexibility the candidate has in selecting appropriate structures and using different structural forms where appropriate, and how repetitive the structures are;

[4] Short, independent sentences such as active structures, simple tense forms (present, past, future), prepositional adjectives, zero and first conditional structures, relative clauses, simple modality (can, must, have to), simple passive voice, question forms (including wh questions) etc.

[5] Longer sentences with subordinate clauses including structures such as a variety of tense grammar including the perfect and continuous aspect, hypothetical conditionals, modals expressing possibility and probability, reported speech, infinitives and gerunds, perfect and continuous passives, etc.

- Error and the extent to which error impedes raters' understanding of the candidate; and
- The extent and success of candidate self-correction.

5.4.4 Vocabulary

Vocabulary focuses on the range of words that the candidate uses and how accurate and precise the candidate's words are. Rating vocabulary involves identifying:

- How well the candidate's lexical resource allows them to address general and technical topics;
- How much lexical range the candidate has and how repetitive the candidate's vocabulary is;
- The extent of precision of meaning, particularly with regard to the use of lower frequency vocabulary;
- The frequency of error in terms of word choice or formulation;
- How a candidate deals with a lack of vocabulary, and how successful circumlocution is (if used); and
- The occurrence and appropriacy of idiomatic language and collocation.

The test tasks, in particular, tasks 1 and 2, are designed to elicit 'foundation' vocabulary, or the generic technical vocabulary that commonly appears across a range of STEM subjects[6]. Candidates are not expected to produce aviation-specific technical vocabulary, though some candidates may have existing subject matter knowledge and associated lexis. Here are some examples of the generic technical vocabulary that Checkpoint candidates use:

- Verbs: Rotate, reach, rise, ascend, deviate, heat, operate, transmit
- Nouns: Phenomenon, temperature, scenario, pressure, markings, instruments, authorities
- Adjectives: Complex, asymmetric, electronic, adverse, rough, corrosive, sophisticated

In task 1, some essential technical vocabulary is presented in the animation to assist the candidate with their description. Candidate's use of this vocabulary should not considered as part of the candidate's lexical resource and is treated instead in the task fulfilment criterion.

5.4.5 Fluency

Fluency focuses on how much language the candidate produces and how smooth and well organised the candidate's language is. Rating fluency involves identifying:

- The speed of the candidate's speech flow or tempo;
- The length of turn the candidate is able to produce, or the ability of the candidate to 'keep going';
- How coherent the candidate is, or how easy their ideas are to follow and understand;
- How effectively the candidate links their ideas using cohesive devices such as discourse markers[7] and grammatical reference[8]; and
- The extent of pausing, hesitation, repetition and self-correction.

[6] See Test reliability and validity document on the Latitude website for a detailed discussion of 'foundation' language

[7] For example: *and, but, however, on the other hand, anyway, the next thing is, and that's it*

[8] For example: There are people surrounding the aircraft *which* (the aircraft) is parked on the stand

5.5 Checkpoint score reports

Test users receive a score report for each test session. Below is a sample test report.

#	FIRST NAME	LAST NAME	LISTENING	READING	SPEAKING	SPEAKING				
						TF	P	S	V	F
1	Wang Qiang	Wong	20.83	33.3	R	R	R	R	R	R
2	Tomas	Schmit	62.50	66.67	Y	G	Y	Y	G	G
3	Rashed	Ali Aish	54.17	47.75	Y	Y	Y	Y	Y	Y
4	Wang Ping	Lui	62.50	70.33	Y	G	Y	G	G	Y
5	Fedhel	Talahi	37.50	41.67	R	R	R	Y	Y	R
6	Claude	Corichon	72.92	82.80	G	G	G	G	G	G
7	Miguel	Serra	45.83	54.17	Y	Y	Y	Y	Y	Y
8	Chanchaio	Chaiprasit	67.50	77.80	G	G	G	G	G	G
9	Nguyen	Ahn Dung	58.33	62.50	Y	Y	Y	Y	Y	Y
10	Maxim	Vakorin	41.67	54.17	R	R	Y	R	R	R

5.6 Using the score report

The test report comprises:

1. Traffic light scores in each of the listening, reading and speaking tests. This allows decision makers to view test performance ‘at-a-glance’. This is particularly useful when selecting or admitting a small number of candidates a large test population. For example, if the purpose were to select two candidates from the ten candidates above based on ELP, a decision maker could quickly identify candidates 6 and 8 as the most proficient.

2. Percentage scores for reading and listening tests and scores by criteria for the speaking test. These more granular-level scores indicate performance within the red, yellow and green levels. These scores are particularly useful when selecting or admitting candidates from a smaller test population where finer distinction between candidate ability needs to be made. For example, if the purpose were to select four candidates from the ten candidates above based on ELP, one might:

- Reject candidates 1, 5 and 10 (as they scored red in some or all parts of the test)
- Select (in order of preference):
  - Candidates 6 and 8 (as they both scored green in all parts of the test)
  - Candidates 2, 4 (as they scored a mix of green and yellow)
  - Candidate 9 in preference to candidates 3 and 7. Although all three candidates scored yellow in all parts of the test, candidate 9’s scores in listening and reading were significantly higher in the level than candidates 3 and 7.