Latitude
Aviation English Services

Checkpoint

LANGUAGE ASSESSMENT FOR STUDENT SELECTION AND ADMISSIONS

Report on the revision of the Checkpoint speaking assessment

July 2022

Contents

1. Introduction

In 2014, following two years of research and development, Latitude launched Checkpoint, a specific-purpose English Language Proficiency (ELP) assessment for student pilot selection and admissions. Development was driven by the following factors:

1. Many aspiring pilots do not have English as a first language and yet undergo ab-initio flight training conducted in the medium of English. The value of ensuring that aspiring pilots enter flight training well equipped, linguistically-speaking, is highlighted by Roberts & Orr (2020, p2): 'Beginning this journey with a safe and effective level of English language proficiency allows for a smooth, cost-efficient, and timely flight training experience'. Conversely, inadequate ELP causes problems for students and their instructors and leads to costly disruption and training failures. At worst, poor ELP is a threat to safety. Therefore, assessment of ELP before a student begins flight training is essential for effective pilot selection and admissions.
2. Whilst there had been much consideration of language assessment for professional pilots and air traffic controllers in response to the ICAO Language Proficiency Requirements (LPRs), we recognised that:
    a. The language skills that students require for successful completion of English-medium aviation training are fundamentally distinct from those required for effective radio communication;
    b. Language tests designed in accordance with ICAO Document 9835 are inappropriate for making decisions about an individual's ability to cope with primary aviation training because:
        i. They assume professional knowledge of aeronautical operations that students do not have;
        ii. They do not measure the skills that students need (for example, interacting with flight instructors and reading training manuals);
    c. General and academic English tests are inappropriate for student selection because:
        i. They do not address the context of aviation training; and
        ii. They test content and skills which are irrelevant to the needs of student pilots.
3. Flight training managers, students and sponsors have a need for a professionally-produced language assessment tool that addresses the specific Target Language Use (TLU) domain of ab-initio aviation training.

To begin with, it is important that we define what Checkpoint is and, equally, what it is not:

- Checkpoint is an accurate, easy-to-use assessment of reading, listening and speaking designed specifically for student pilot selection and admissions decisions.
- Checkpoint is designed for use alongside existing skills assessment procedures to determine language proficiency for entry to ab-initio flight training programmes and to identify any language training requirements.
- Checkpoint is web-based for reasons of flexibility, practicality and cost-effectiveness. The stakeholders - airlines, flight schools, colleges and universities - are responsible for administration of the assessment.

- In keeping with the ICAO LPRs, Checkpoint is not intended to be used for the assessment of language for radio communications for personnel licensing purposes.
- Checkpoint operates according to specific-purpose assessment criteria including a discrete rating scale for the assessment of speaking. Resulting from targeted needs analysis, the Checkpoint rating scale descriptors are designed specifically for the stated purpose and do not correspond to the descriptors found in the ICAO Rating Scale.
- Checkpoint scores are designed to align with the Common European Framework of Reference for Languages (CEFR)[1] and are reported using a traffic light system as follows:

🔴 RED - CEFR B1 and below: Language is likely to be an obstacle to successful aviation training for candidates that score red in any part of the assessment. We recommend that candidates who score red in any part of the assessment undergo at least 200 hours of language training before beginning flight training.

🟡 YELLOW - CEFR B1-B2: Candidates that score yellow in any part of the assessment may encounter language-related difficulties during aviation training. We recommend that candidates who score yellow in any part of the assessment undergo between 25-200 hours of language training before beginning flight training.

🟢 GREEN - CEFR B2 and above: Candidates that score green in all parts of the assessment are unlikely to encounter language-related difficulties during flight training.

To date, Checkpoint has helped numerous organisations in their selection and admissions process by assessing over 2,000 candidates providing not only guidance on ELP but also identifying areas of weakness so that candidates can improve their English language skills in a targeted and efficient manner.

[1] See: https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions.

2. Rationale and timeline for the revision of the speaking assessment

In line with prevailing attitudes in the language testing community, language assessments have a limited life span. Use of an assessment over time may reveal signs of a reduction in its capability to fulfil its initial function effectively which might manifest in the form of criticism, developments in the TLU domain or general observations (Alderson *et. al.*, 1995, p227). The revision timeline below demonstrates how the rationale for a revision of the Checkpoint speaking assessment emerged from a concatenation of events. These individual signs, when taken as a whole, suggested that revision was required to ensure the speaking assessment's fitness for purpose.

REVISION TIMELINE

June 2020:

- Observations from rater training and harmonisation meetings collected.
- Plan to conduct needs analysis and revise the speaking assessment drawn up.

26th October 2020: Publication of an article that reflects on the strengths and weaknesses of Checkpoint (Lynch & Porcellato, 2020).

4th November 2020: Latitude responds to Lynch & Porcellato.[2]

11th March 2021: Formal data gathering activities begin.

27th May 2021: Proposal submitted for UK Economic and Social Research Council (ESRC) funded project in collaboration with Professor Tineke Brunfaut and Dr Olena Rossi, Lancaster University, UK.

21st June 2021: ESRC proposal accepted.[3]

8th September 2021: Lancaster delivers ESRC research report to Latitude.

27th September 2021: Finalisation of schedule for the implementation of report recommendations.

October 2021:

- Finalisation of new task specifications.
- Drafting of new task content.

November 2021:

- Revision of Checkpoint rating scale criteria and descriptors begins.
- Pilot trials begin.

1st December 2021: Revision and finalisation of task and item modifications.

December 2020 - March 2022:

- Secondary trials conducted.
- Production of updated candidate familiarisation videos.

February - March 2022: Standard-setting and rater training.

11th March 2022: Checkpoint speaking version two launched.

[2] See https://revistas.pucsp.br/index.php/esp/article/view/51379/33565.

[3] View the press release here.

3. Research and needs analysis

Research began internally in June 2020. The project opened with a review of the literature on oral language assessment, English for specific purposes, integrated listening-speaking tasks, ab-initio flight training and research methods in applied linguistics, notably on conducting semi-structured interviews and designing questionnaires. To further investigate ELP in the domain of flight training, a questionnaire targeting the opinions and experiences of student pilots at flight school was prepared and distributed electronically. We also began approaching flight training organisations in order to interview flight instructors concerning their perceptions of the language skills needed by their students for successful participation in and ultimately completion of flight training.

4. Research with Lancaster University

In late May 2021 we received news from a post-doctoral researcher at Lancaster University regarding an opportunity for UK Economic and Social Research Council (ESRC) funded research into the Checkpoint speaking assessment. The funding would come in the form of *Accelerating Business Collaboration*, an initiative designed to enable postgraduate and early-career researchers in the social sciences to successfully collaborate with non-higher education partners through research. On 21st June 2021, the Department of Linguistics and English Language at Lancaster University, in partnership with Latitude, won an *Accelerating Business Collaboration* award to research spoken English for successful flight training. Latitude already had a long-standing partnership with Lancaster University and we were pleased to collaborate with Professor Tineke Brunfaut and Dr Olena Rossi to support the study. We delivered data obtained through our internal research, literature, aviation training documentation, a list of potential flight instructor informants and a selection of Checkpoint speaking samples to Lancaster, and the research was carried out by Professor Tineke Brunfaut and Dr Olena Rossi in July and August 2021. On the 8th September 2021 we received a comprehensive report from Lancaster that combined the researcher's findings with recommendations on how to address the issues that their research had identified. Latitude then assumed responsibility for the revision of the Checkpoint speaking assessment including implementation of the report's findings.

5. Areas addressed by the revision

5.1 Specifications

One of the issues that we had identified with the existing assessment tasks and their sequencing within the speaking assessment as a whole was that, in their original form, ensuring comparability between versions was complicated. Following the recommendations in Lancaster's report and referring to the literature on developing test specifications, we re-drafted individual item specifications and used these as the blueprint for the revised Checkpoint speaking assessment and subsequent versions. Each item has an individual specification comprising *general description*, *prompt attributes*, *response attributes* and a *sample item*. A specification supplement provides *extra guidance and information* which may be useful for future development. The *prompt attributes* and *response attributes* may overlap but essentially the first details what will be given to the candidate and the latter 'describes what should happen when the test taker responds' (Davidson & Lynch, 2002, p25). By providing detailed prompt and response characteristics we are endeavouring to ensure comparability between different versions of the speaking assessment. In keeping with guidance in the literature, we followed an iterative specification drafting process throughout the entire development procedure to ensure "item/task fit-to-spec" and accurate guidance for future item writers and assessment developers. The result is four parallel sets of tasks based on identical requirements and a rigorously-applied construct.

5.2 Situational language use

It's important to remember at this stage that the spoken language proficiency of student pilots is challenged in a number of situations at flight school. These are detailed below but they fall broadly into the following categories:

1. Inside flight training:
    a. Ground training;
    b. Flight training; and
2. Outside of flight training.

The objective of Checkpoint is to assess an individual's capacity to cope on arrival in a flight training environment where the lingua franca is English. Accordingly, a speaking assessment instrument needs to evaluate whether someone can express themselves spontaneously and clearly in English in the situations above using language strategies to overcome any shortcomings in technical or professional knowledge. As assessment developers it is our responsibility to set a level playing field for candidates particularly as Checkpoint is often used for high-stakes assessment for selection for airline-sponsored flight training programmes, successful completion of which leads to secure employment.

5.3 Subject matter content

If we are to make judgements about a candidate's ability to cope with flight training, inclusion of domain-related content is essential for construct validity, in other words, the extent to which a test measures what it claims to measure. However, as Checkpoint candidates are applicants for ab-initio flight training, we have to account for the fact that many have zero flight hours and little knowledge of aviation. This presents a challenge for assessment developers. On one hand, we need to include aviation-related content for the purposes of construct validity and to satisfy user expectations of what an "aviation English" assessment might look like, and to control content and tasks to ensure that measurement of ELP is independent of candidate's subject matter knowledge on the other.

In the case of assessing language proficiency for student pilot selection and admissions, controlling specific-purpose content is more challenging in assessments of productive skills than in assessments of receptive skills. Listening and reading may be assessed legitimately by the inclusion of expository texts of the sort typically found at flight school. Authentic extracts from theoretical training courseware and ground school classes which introduce and explain essential topics such as the navigation or air law are highly representative of the type of language-use tasks that student pilots face during flight training. Expository listening and reading texts, by their very nature, are created to be accessible to the layperson and, if selected correctly, should provide all the information necessary for reader or listener to "learn" about the topic in a manner similar to that which students acquire technical knowledge during training. Provided that items are carefully drafted to tap language knowledge specifically, in other words, answering questions correctly is contingent on the candidate's knowledge of language alone, then item responses can be taken as a valid measure of the candidate's ability to understand language as it appears in the domain.

Assessing speaking differs from assessing listening and reading in that prompts need to elicit a spoken performance from the candidate. Constructing a spoken performance involves expressing thoughts, ideas and opinions which requires drawing on not only language knowledge, but knowledge of the world and specifically the subject matter presented in the prompts themselves. Given the layperson candidate, there is the potential for speaking prompts with strong aviation themes to disadvantage candidates who do not possess knowledge of aviation on one hand, and be biased in favour of those who have prior knowledge or experience on the other.

In the original version of the Checkpoint speaking assessment, task 1 featured a visual representation of an aviation-related process or mechanism. Task version content included topics such as the jet engine, primary and secondary radar and, in meteorology, the Foehn effect. These prompts were presented as either an animation or a series of still images with low-frequency technical vocabulary embedded in the prompt to support the layperson candidate. Time was given for the candidate to study the prompt before producing an oral description. Despite feedback we received from candidates that indicated that the task was strongly representative of the sort of speaking tasks that students have to perform at flight school, our rater team observed that some candidates produced performances indicative of acquired aviation knowledge or previous flying experience which raised a concern about the potential for conflation of language and subject matter knowledge in language assessment. In our revision of the Checkpoint speaking assessment, we were mindful of this potential for measurement error. In accordance with guidance from Lancaster, we took the decision to withdraw this task and develop a replacement along with the specification that all speaking tasks and prompts should be relevant to the aviation training domain for the purpose of construct validity but should not be of a technical, aviation-specific nature.

5.4 Communicative functions

Based on semi-structured interviews with seven flight instructors working in FAA and EASA flight training programmes, the research team at Lancaster produced a taxonomy of communicative functions (figure 1) for the TLU domain of ab-initio flight training which we used as the basis for the development of revised task and item specifications. The taxonomy was instrumental both in our review of the existing tasks and in the drafting of a replacement task designed to target the CEFR B2-C1 communicative functions that correspond to 'green' on the Checkpoint rating scale.

Our initial in-house research findings had pointed towards an integrated-skills item, either a reading-speaking or a listening-speaking task. Lancaster's report cemented this idea with a recommendation that we develop such a task to allow for the elicitation of higher-level communicative functions such as *synthesising and evaluating*, and even those that are traditionally more difficult to extract using a computer-administered test such as *agreeing and disagreeing*. The evolution of the new task is detailed below but was designed to probe deeper and elicit a wider range of communicative functions identified from the TLU domain to ensure construct validity.

Training stage	CEFR level	Communicative functions
I. Inside of flight training		
1. Ground training	B1	Managing interaction (e.g. to ask a question, to offer own idea)
	B1	Asking for clarification
	A2/B1	Describing places (flight geography)
	A2	Describing things (e.g. weather conditions, plane parts)
	A2	Describing routine operations
	A2	Talking about obligation and necessity
2. Flight training		
Pre-flight briefing	B2	Describing hopes and plans
	B2	Speculating about future events
	B1/B2	Expressing / asking for opinion
	B1/B2	Agreeing / disagreeing (with the instructor about the flight plan)
	B1/B2	Checking understanding (of the flight plan)
Flight practice	B2	Giving precise information (callouts, repeating ATC tower commands)
	B1/B2	Checking understanding (of instructions by the instructor and the ATC tower)
Post-flight briefing	C1	Expressing opinions tentatively, hedging
	C1	Speculating and hypothesising about causes and consequences
	B2/C1	Synthesising and evaluating
	B2/C1	Critiquing and reviewing
	B2	Describing past experiences
	B2	Expressing abstract ideas (e.g. ethics of pilot's actions)
	B2	Expressing certainty, probability and doubt
	B1/B2	Agreeing / disagreeing
	B1/B2	Describing feelings, emotions and attitudes
	B1	Asking for clarification
	A2	Giving suggestions
II. Outside of flight training		
	A1	Giving personal information
	A1	Asking for directions
	A1	Understanding and using prices
	A1/A2	Describing habits and routine
	A1	Telling the time
	A1	Greetings and goodbyes

Figure 1: Communicative functions in ab-initio flight training

5.5 Discrimination between higher and lower-level language users

Checkpoint was developed initially with reference to a needs analysis carried out primarily at flight schools providing EASA flight training programmes. FAA and EASA deviate in that whilst the latter devotes a comparatively long period of time to initial theoretical knowledge instruction, students on FAA programmes begin practical flying training much earlier in their training programme. Consequently, they are more likely to require specific communicative functions at an earlier stage compared to their counterparts on EASA programmes. Indeed, Lancaster's report suggested that whilst the communicative functions observed during ground school ranged from CEFR A2-B1, during practical flying training students would need to employ language typical of CEFR B1-C1. As Checkpoint has become more widely used for selection and admissions to FAA Part 61 and Part 141 flight training programmes, the speaking assessment tasks required revision in order to elicit both lower and higher-level language functions to determine a candidate's readiness for the linguistic demands of practical flying training.

Lancaster's research explored the lexical and structural complexity of 45 speech samples generated from the first version of the Checkpoint speaking assessment. Lexical complexity was analysed in terms of:

- **Lexical density** measured by calculating the proportion of content words (nouns, verbs, adjectives and adverbs) to the total number of words in the samples;
- **Lexical diversity** measured by calculating the type-token ratio, or the number of unique words in relation to the total number of individual words in samples; and
- **Lexical sophistication** measured by calculating the percentage of words that are among the 500, 1000 and 2500 most common English words as defined by the New General Service List (Brezina & Gablasova, 2015).

In terms of structural complexity, calculations were made of sample length providing data for total length, mean sample length and range in each band and for each individual task. An analysis of the length of production measured the length of clauses in samples and the length of T-units[4] across the existing Checkpoint versions. A more in-depth analysis of the number of clauses per T-unit and the number of simple clauses per T-unit allowed the researchers to draw conclusions about the utterance complexity for each band level. Similar methodology was used to analyse subordination and coordination measuring respectively the proportion of subordinate and coordinate clauses to the overall number of clauses and T-units. Cohesion was measured by grouping connectors into categories such as time sequence, cause and effect, persuasion and emphasis. These categories were then assigned a CEFR proficiency level based on research by North *et. al.* (2010) and examples of the connectors were searched for in each of the green, yellow and red bands referring to relevant discourse markers and functions per CEFR level.

The findings of Lancaster's research substantiated observations made by our rater team that the tasks in their current form could be enhanced to elicit a more substantial and varied sample of speech from candidates. The in-depth quantitative and qualitative analyses indicated that the length and quality of the language samples supplied

[4] A T-unit is defined as 'an independent clause and any clauses dependent on it' (Hunt, 1965).

sometimes made it difficult to discriminate in borderline cases. It was also noted that improvements could be made in the consistency of datasets from the four different speaking task sets used for the Checkpoint speaking assessment.

The quantitative analysis was reinforced by qualitative observations of individual tasks. Equally, feedback from our rater team confirmed that candidate speech sometimes contained insufficient evidence of cohesive devices, discourse markers or subordinate clauses that would help distinguish more proficient speakers from lower-level users of English. In accordance with the advice provided by Lancaster, we began work to restructure the test and tasks to address these concerns.

6. Restructuring the speaking assessment

We began by rearranging the speaking assessment into three tasks as follows:

6.1 Speaking task 1

Task 1 comprises a set of three questions designed to evolve in terms of the complexity of the response required.

EXAMPLE TASK PROMPTS:

QUESTION 1: *Why do you need to speak English to be a pilot or an air traffic controller?*

QUESTION 2: *Why do you think pilots are well paid? Should they earn more money than train or bus drivers? Why? Why not?*

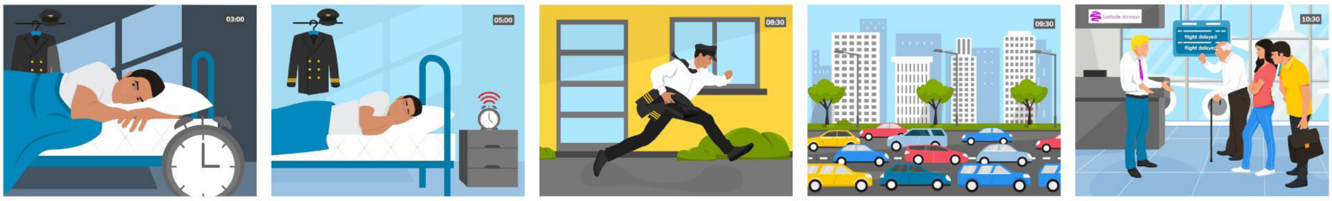
QUESTION 3: *How does building an international airport affect a country? What are the advantages and disadvantages for the people of that country?*

This task was moved from its original position at the end of the speaking assessment to the beginning. Conscious of the limitations associated with web-based speaking tests, we carried out this change as we felt that the clarity and simplicity of question-answer routines would help candidates feel more at ease at the start of the assessment. In addition, the prompts have been constructed in such a way that they elicit communicative functions ranging from A2 to C1 on the CEFR scale by means of the initial question and, where appropriate, an auxiliary prompt. Accordingly, most candidates, regardless of their ELP level, should be able to offer a response while the prompts help ensure that those with higher-level language skills have sufficient opportunity to demonstrate their ELP effectively. The available response time for each item was extended from 40 to 90 seconds giving candidates ample opportunity to respond in a full and detailed way. In the interest of test-taker motivation, the content of the items is related to aviation but is designed such that a successful response does not require any technical aviation knowledge on the part of the candidate.

6.2 Speaking task 2

The storyboard narration task was retained but with an enhanced task specification and prompts to afford the candidate greater opportunity to demonstrate their ELP. The original storyboards were scrutinised by our in-house subject-matter experts in the light of our concern around the inclusion of technical aviation content. This led to the development of a new specification for the storyboards to contain topics that pertain to aviation yet remain accessible and easy to understand for the layperson candidate. Artwork briefs were created and then the storyboards were produced by a professional illustrator. The illustrations were the subject of several drafts and revisions before a set of storyboards, both homogenous and unambiguous, was finalised. The original response time for this task was preserved at 90 seconds.

EXAMPLE TASK PROMPT:



One limitation of the original storyboard narration task was that some candidates, even those with higher levels of ELP, produce responses limited to linear sequences of descriptions with basic linking devices and a more restricted structural range. Therefore, following Lancaster's recommendations, we developed three follow-up questions for each storyboard to:

- Prompt candidates to develop their responses related to the themes raised in the storyboard; and
- Elicit a longer and more varied sample of speech.

This sub-task also provided the opportunity for the test designers to tap the ability of candidates to produce question forms, a language skill that is of high importance in flight training as students engage in 'a steady stream of interrogation' with their instructors (Udell & Schneider, 2021). The response time for the three follow-up questions is two minutes.

EXAMPLE TASK PROMPTS:

QUESTION 1: *What factors might cause a commercial pilot to be tired?*

QUESTION 2: *If the pilot had woken when the alarm clock rang, how could the story have been different?*

QUESTION 3: *Imagine you are the pilot's employer. Ask the pilot three questions about the situation.*

In the examples above, the candidate produces a narration of images depicting a pilot who has difficulty sleeping and consequently sleeps through his alarm then is late for his flight. The three follow-up questions develop the theme of fatigue. Although this is a pertinent topic of concern in the aviation industry, the content is general enough for the layperson to offer a response.

6.3 Speaking task 3

6.3.1 Initial considerations

Our decision to withdraw what was task 1 in the original speaking assessment presented an opportunity to strengthen the relationship between assessment tasks and those tasks that candidates would be likely to perform in the real world. Where tasks 1 and 2 are designed to elicit spoken performances which are to a greater extent independent of other skills, for task 3, we elected to design an integrated skills task. In essence, an integrated skills task is one that requires a candidate to 'integrate multiple language skills in a substantial way to complete a speaking task' (Lee, 2015). This could include, for example, listening to a lecture or presentation and subsequently giving an oral account of what was heard. Alternatively, the task prompt might be in written form where the test taker reads a text before summarising or giving an opinion on the subject matter. On consulting the literature, we felt an integrated skills task would add considerable value to the Checkpoint speaking assessment for a number of reasons:

- 'Many real-world communicative acts rely on the integration of two or more … skills as well as other non-linguistic cognitive abilities' (Frost, 2010, p346);
- Much spoken performance at flight school:
  - Is 'goal-oriented' and 'situated in specific settings' (Bachman, 2002);
  - Is contingent on the skills of listening, for example, to instructors and other students, and reading, for example, technical documents and training manuals;
  - Requires students to 'speak coherently about relevant ideas, to handle source documents appropriately and to display knowledge in relevant ways' (Cumming, 2014, p1);
- 'Integrated tasks are authentic because they provide a realistic context for speaking performance and require test-takers to perform tasks that are relevant' (Barkaoui *et. al.*, 2012, p2); and
- Integrated tasks seem to 'offer a better means of assessing the traits underlying language proficiency than is possible if the skills are tested in relative isolation from each other' (Weir, 1990, p84).

6.3.2 Development steps

Having elected to develop an integrated skills task, the initial challenge was to decide on the form and content of the input text. Following guidance from McNamara (1996), authentic training texts on topics such as principles of flight or meteorology were dismissed because, although highly representative of the TLU domain, it was not considered appropriate to prompt layperson candidates to construct spoken performances in relation to strong technical subject matter content. To provide candidates with an equal opportunity to perform to the best of their ability regardless of prior knowledge, we were mindful of the need for input texts to be accessible to the layperson candidate while eliciting the higher-level communicative functions required of students at the different phases of flight training as identified in Lancaster's report, for example, *expressing opinions* (B1/B2), *giving precise information* (B2) *expressing certainty, probability and doubt* (B2) *synthesising and evaluating* (B2/C1) and *synthesising and hypothesising* about causes and consequences (C1). Further considerations were the need to present texts that are representative of scenarios that ab-initio pilots might encounter at flight school for reasons of authenticity, and to tap speaking skills required both inside and outside of flight training itself. Indeed, research by Bieswanger *et. al.* (2020) reported the difficulties that some Non-Native English Speakers (NNESs) encounter when using English as a lingua franca for daily routines such as speaking about accommodation, ordering food and dealing with flight school administration.

Due to the co-dependency of listening and speaking skills in many flight training situations, and taking into account the considerations above alongside the need to produce several versions of equivalent content, we developed a specification for a listening-speaking task featuring a pre-recorded oral presentation of approximately four minutes delivered by a flight school instructor to a class of student pilots. With guidance from our in-house subject matter experts, we developed four plausible storylines and scripted dialogues for the instructor and two or more student interlocutors. Care was taken to include US and British English accents, male and female voices and NNESs in the production of the recordings. Informal trials with native-speaker participants led to the decision to reduce the speech rate for the presentation to between 140-160 words per minute.

Conscious of the cognitive burden on the candidate, we choose to include visual prompts as part of the input text. We created a presentation of five slides to accompany each text whose purpose is to clarify obscure information such as place names, people's titles and names or reference to unusual equipment, as well as providing visual support for lower-level candidates. The presentations run automatically on the screen alongside the spoken input.

The rubrics were set out to inform candidates how the task progresses from listening and note-taking to planning and execution using structured prompts to clarify the task requirements and to encourage candidates to fully develop their responses. To avoid initial concerns that candidate responses could represent simple recount of input rather than broader interaction with the text, candidates are advised to include:

- A summary of the topic;
- Specific details that they think are important and why; and
- Their opinions about the planned activity and the arrangements.

6.4 Length of overall speech production

A further desirable result of expanding the range of targeted communicative functions in the Checkpoint speaking assessment is that we were able to increase the available response time. With the modification of each task we have succeeded in increasing the available response time from up to 4.5 minutes of rateable speech in the first version to up to 11 minutes in the revised version. Results from trials indicate that at least 10 minutes of rateable speech is typically generated. It should be noted that tasks 2 and 3 also incorporate planning time of 60 seconds per task.

The enhanced tasks also led to an increase in administration time from approximately 10 minutes in the first version to approximately 25 minutes in the revised version. For this reason, we took the decision to separate the speaking assessment from the listening and reading assessments so that it stands alone in the assessment platform. This enables candidates to take a break between assessments, thus optimising conditions for best performance.

7. Trialling

7.1 Pilot trial

In November 2021, we ran a pilot trial of the blueprint for the speaking assessment with seven Arabic, Polish, Portuguese and Vietnamese speakers of English. The candidates were all aspiring pilots at different stages of their flight training programmes ranging from those who had not yet started to others who had completed some flight training. The assessment featured a post-assessment feedback form completed by candidates. The pilot exposed some weaknesses in the assessment structure, rubrics and individual tasks and enabled improvements to be made to the task specifications, including the major redesign of the task 2 storyboard before moving to a secondary trial.

7.2 Secondary trial

In early 2022 we engaged with three training organisations to trial the revised speaking assessment alongside the existing listening and reading assessments in close-to-real assessment contexts. 28 Korean, French and Vietnamese speakers of English participated in the secondary trial. As with the pilot trial, all candidates were aspiring pilots at different stages of their flight training programmes.

The speech samples generated from the trials were used by the test developers:

- To verify the assessment functionality, structure and timing and clarity of rubrics and prompts;
- To verify that the tasks were generating an adequate sample of speech representative of the target language detailed in the specifications;
- As the basis for revision of the rating scale for the Checkpoint speaking assessment; and
- As a resource for internal standard-setting and rater training.

8. Rating scale revision

As with the majority of speaking assessments, the Checkpoint speaking assessment is marked by trained human raters with reference to an analytic rating scale. The modification of existing tasks and the inclusion of new ones necessitated a revision of the criteria and descriptors in the existing rating scale to reflect the changes and to more fully capture the language requirements at the CEFR B2/C1 levels as identified in Lancaster's report. The changes were drafted and then revised and finalised with reference to the speech samples gathered from trials. These changes are summarised below.

8.1 Task fulfilment

Each band and task descriptor has been updated to reflect the new task characteristics. Where appropriate we incorporated descriptors for thematic development found in the CEFR Companion Volume (Council of Europe, 2018, p141) thus ensuring that the bands are consistent with the range of proficiency from A2 to C1. Particular attention was paid to the descriptors relating to the integrated listening-speaking task (task 3). In their research into integrated listening-speaking tasks, Frost *et. al.* (2011, p336) reported that 'the number of accurate replications of ideas from the source text increased according to proficiency and the number of distortions of the source text information declined as proficiency level increased'. Therefore, we incorporated specific descriptors to account for the variations in the accuracy of content displayed in candidate performances at the three levels.

8.2 Pronunciation

Following the advice from Lancaster, we removed any reference to technical vocabulary.

8.3 Language-in-use

A growing body of research is analysing and questioning the practicality of trying to separate lexis and grammar in language assessment. Referring to research by Römer (2017), we took the decision to merge the distinct criteria of *vocabulary* and *structure* into one new criterion, *language-in-use*, enabling clearer and more practical guidance to our rater team. In addition, as with pronunciation, we removed any reference to technical vocabulary.

8.4 Fluency

The original *fluency* criterion considered both fluency and cohesion and coherence in a single criterion. Following advice from Lancaster that these aspects of language performance are distinctly different and that merging them might lead to misunderstandings, we took the decision to separate these aspects of spoken performance to form two separate criteria. The fluency criterion was improved with reference to rater feedback gathered during previous standard setting meetings and the spoken fluency band descriptors contained in the CEFR Companion Volume (Council of Europe, 2018, p144). Any references to coherence and cohesion descriptors were removed, clarifying a focus on features of performance relating exclusively to speech production and flow. We have endeavoured to draft this criterion so that it distinguishes better between performances from band to band and provides clearer, more practical guidance for raters.

8.5 Coherence and cohesion

The addition of this final criterion has allowed us to consider with greater focus how well a candidate organises speech and the extent to which the language is connected and logical. An advantage of the *coherence and cohesion* criterion is that it helps the rater distinguish more easily between higher and lower-level performances as it allows descriptions, based on evidence from trial speech samples, of the nature and variety of cohesive devices and the organisational features characteristic of each band level.

9. Standard setting and rater training

9.1 Standard setting

After the secondary trial was complete and the revised rating scale drafted, the test development team held a series of standard-setting meetings consisting of table-top discussions orientated to linking features of candidate performance to the revised scale. The trial speech samples, both in audio format and with transcriptions, were interrogated with reference to the draft of the revised scale to identify:

- Exemplar performances at the red, yellow and green levels which could be used as 'benchmark' performances for team standardisation;
- Further performances suitable for the purposes of individual rater training and certification; and
- Any further revisions to the rating scale that were required.

9.2 Rater training

Prior to launch, all Checkpoint raters underwent thorough re-training following the steps outlined below.

9.2.1 Pre-standardisation

An initial meeting was held during which the test developers explained the rationale behind the revision and provided an outline of the major changes to the specifications and tasks. Each rater then familiarised themself more deeply with the test format by watching the test familiarisation videos and listening to one exemplar performance at each of the red, yellow and green levels.

9.2.2 Standardisation

We held a second rating team meeting to:

- Discuss the interpretation of the rating scale and procedures for rating each of the three speaking tasks; and
- Listen to, rate and discuss a series of exemplar performances at the red, yellow and green levels.

9.2.3 Certification

Finally, each rater rated a set of speech samples independently, returning scores to the team leader. The team leader analysed the rating data with many-facet Rasch analysis using the computer program FACETS[5] and returned feedback on consistency and severity to the individual raters.

We aim to achieve infit mean square values of between 0.50 and 1.50 and, where applicable, outfit mean square values of between -2 and +2 (Green, 2013, pp219). At the point of certification, our current rater team operated with in-fit values of between 0.61 and 1.34.

In cases of disagreement we met to discuss the data and the individual raters' justifications for the scores awarded which led to an overall harmonisation of scores and greater shared understanding of the rating scale and therefore inter-rater reliability. This process also yielded a further opportunity to draft guidance notes to accompany the revised rating scale to assist raters in making their assessment decisions.

[5] Linacre, J. M. (2010). FACETS Rasch measurement computer program. Chicago: Winsteps.com

In-line with best practice, we will continue to analyse rater performance at periodic intervals during testing cycles. We will do this by:

- Asking raters to rate ‘reliability samples’ for the purposes of calculating intra- and inter-rater reliability; and
- Selecting rated samples at random for second-rating by the team leader, with discrepancies discussed with the rater in question.

10. Familiarisation videos

Before the official launch of the revised speaking assessment we published a series of informative videos to assist potential Checkpoint candidates in understanding the task requirements and familiarise themselves with the assessment procedures. You can view the videos here: https://www.latitude-aes.aero/checkpoint-familiarisation.

11. Revision summary

The revision of the Checkpoint speaking assessment has resulted in the following enhancements:

- An increase of available response time from 4.5 to 11 minutes;
- A new integrated listening-speaking task engaging higher-level language skills;
- Tasks capable of eliciting communicative functions typically required at flight school from CEFR A2 to C1;
- Removal of technical subject matter content from task prompts;
- A revised rating scale that:
    - Is more closely aligned with the CEFR;
    - Better reflects the linguistic features of candidate performance at each of the three levels;
- Revised and more detailed task specifications to ensure greater equivalence between versions;
- Updated, clear and informative familiarisation videos to enable candidates to prepare;
- Optimised conditions for candidates to perform at their best:
    - Separation of speaking from listening and reading to enable a break in assessment; and
    - Re-worked assessment structure, format and rubrics to facilitate candidate engagement.

12. Future directions

In line with our commitment to providing quality aviation English assessments, we remain open to collaboration with industry partners and researchers in applied linguistics so we can better understand how Checkpoint functions so that we can provide the best possible service to our customers and candidates. With regard to the revised Checkpoint speaking assessment, we are particularly interested in investigating:

- The lexical and structural complexity of candidate performances generated from the revised Checkpoint speaking assessment;
- The correspondence between features of candidate performance and the Checkpoint rating scale; and
- The alignment of the Checkpoint rating scale with the CEFR.

The purpose of Checkpoint is to help flight schools and aspiring pilots achieve flight training success, and naturally, our primary interest is in research that helps us to understand if Checkpoint is achieving this objective. If you are interested in collaborating with Latitude and would like to find out more, contact us at checkpoint@latitude-aes.aero.

13. References

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

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, 19(4), 453–476.

Barkaoui, K., Brooks, L., Swain, M. & Lapkin, S. (2012). Test-Takers' Strategic Behaviors in Independent and Integrated Speaking Tasks. *Applied Linguistics* 2012: 1–22.

Bieswanger, M., Prado, M. & Roberts, J. (2020). Pilot training and English as a lingua franca: some implications for the design of Aviation English for ab initio flight training courses. *The ESPecialist* 41(4).

Brezina, V. & Gablasova, D. (2015). Is There a Core General Vocabulary? Introducing the "New General Service List". *Applied Linguistics* 36, 1-22.

Council of Europe (2018). Common European Framework of Reference for Languages: Learning, Teaching, Assessment - Companion Volume with New Descriptors.

Cumming, A. (2014). Assessing Integrated Skills. in Kunnan, A. J. (ed.) *The Companion to Language Assessment.* John Wiley & Sons.

Davidson, F. & Lynch, B. K. (2002). Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications. Yale University.

Frost, K. (2010). Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test takers' oral performances. Masters Dissertation, University of Melbourne.

Frost, K., Elder, C., & Wigglesworth, G. (2011). Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test takers' oral performances. *Language Testing* 29(3) pp 345–369.

Green, R. (2013). Statistical analyses for language testers. Palgrave Macmillian.

Hunt, K. (1965). Differences in Grammatical Structures Written at Three Grade Levels. Champaign, IL.

Lee, H.W. (2015). Innovative assessment tasks for academic English proficiency: an integrated listening-speaking task vs. a multimedia mediated speaking task. Graduate Theses and Dissertations 14584 https://lib.dr.iastate.edu/etd/14584

Lynch, J. K., & Porcellato, A. M. (2020). The Case for an Aviation English Screening Tool for US Flight Schools. *The ESPecialist*, 41(4).

McNamara, T. (1996) Measuring second language performance. Longman.

North, B., Ortega, A. & Sheehan, S. (2010). British Council – EAQUALS core inventory for general English. British Council / EAQUALS.

Roberts, J., & Orr, A. (2020). Language Education for Ab Initio Flight Training: A Plan Going Forward. Engaging the Next Generation of Aviation Professionals. https://doi.org/10.4324/9780429287732

Römer, U. (2017). Language assessment and the inseparability of lexis and grammar: Focus on the construct of speaking. *Language Testing* 34(4) pp 477-492.

Schneider, A. & Udell, R. (2021). A corpus-driven approach to Aviation English in pilot flight training. Paper presented at the 8th Grupo de Estudos em Inglês Aeronáutico Seminar, Tuesday 9th November 2021.

Weir, C. J. (1990). Communicative Language Testing. Prentice Hall.

14. Acknowledgements

We would like to extend our sincere thanks and gratitude to Professor Tineke Brunfaut, Dr Olena Rossi, Camilo Ramos and Charlotte Stuart at Lancaster University, and to the UK Economic and Social Research Council.

We would also like to thank the following people for their kind support in helping us to gather data and conduct assessment trials:

- Ahmed AlJowder at Oxfordsaudia Flight Academy, Saudi Arabia.
- Dr Aline Pacheco at Pontifical Catholic University of Rio Grande do Sul, Brazil.
- Alison McCheyne at the École Nationale de l'Aviation Civile, France.
- Mickaël Touaty at the École Supérieure d'Aviation, France.
- Adrian Teo Renlin of TRAVCON Consulting Group.

Last but not least, our thanks to the many student pilots and flight instructors who gave their time so generously for our questionnaires, interviews and assessment trials.