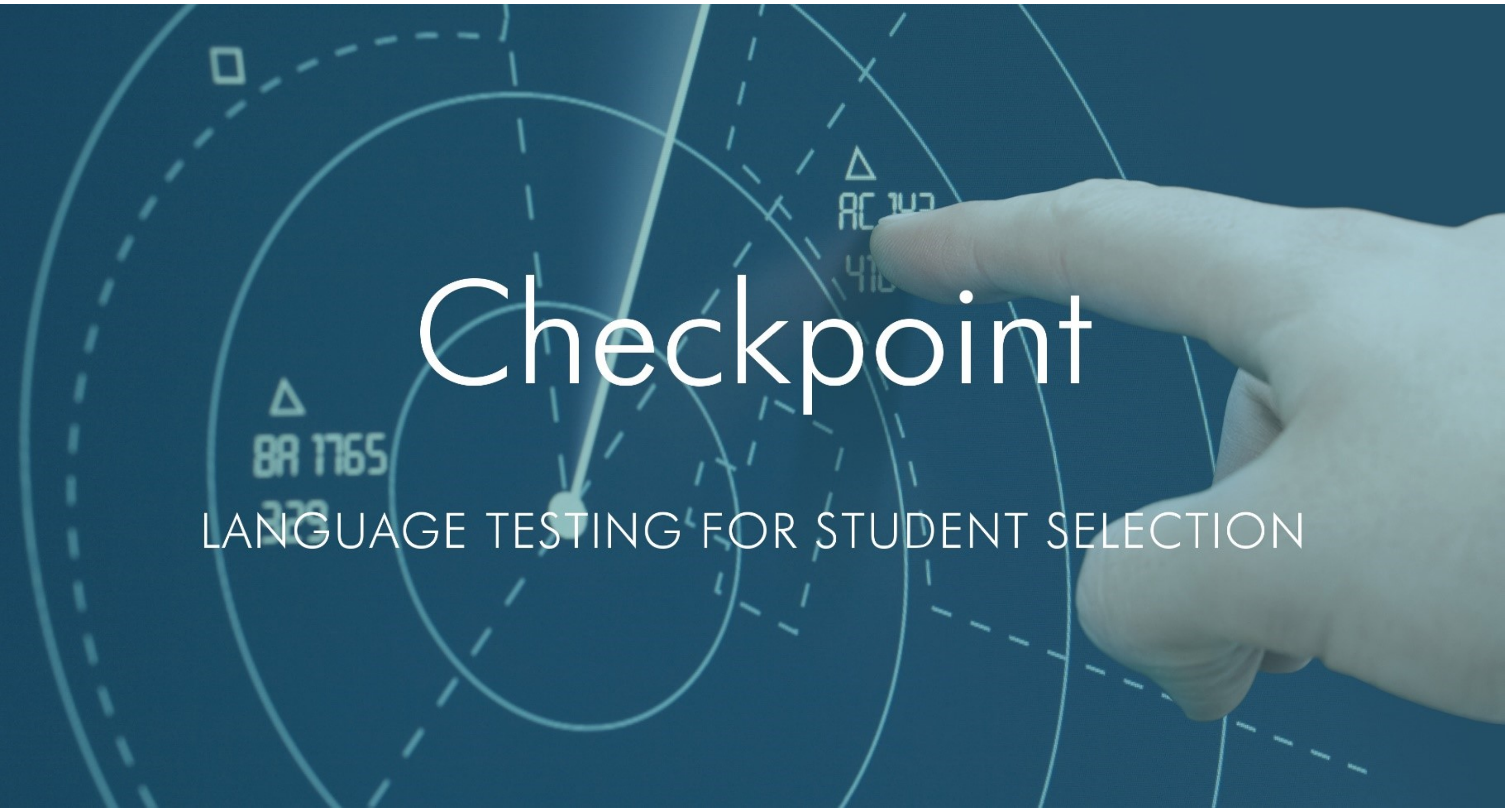
Latitude
Aviation English Services



Checkpoint

LANGUAGE TESTING FOR STUDENT SELECTION

Test reliability and validity

May 2017

Contents

1. Executive summary

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).

Checkpoint has been developed by aviation English testing experts according to the highest standards of language test development. With considerations of validity, reliability and practicality central in the test development process, we have made every effort to develop an instrument which is fit for its intended purpose. This document presents information on the work we have done to establish test reliability and validity and also describes two post-graduate level research investigations into Checkpoint test quality.

To date more than 1,200 candidates have taken versions of the Checkpoint test. Approximately 500 candidates took early versions of the test. Their data was analysed leading to item revisions and improvements and the creation of the two current versions of the test (referred to here as the A-set and the B-set). Two parallel test versions are required to allow for test/re-test of candidates using completely different content. Item difficulty between the two versions has been standardised.

Section 2 of this document provides an introduction to language test validity and reliability. Section 3 presents data on Checkpoint test reliability including Cronbach alpha reliability results, standard error of measurement, listening, reading and speaking test correlations and item discrimination analysis. Currently 702 candidates have taken the final versions of the A-set and B-set tests. Their data show test reliability of >90% for both test versions. To compare Checkpoint's reliability with either formal or informal interviewing, both of which are common methods of assessment for student selection and admissions, Conway, Jako, and Goodman (1995)[1] found that the average reliability of highly structured individual interviews was >59%, while the reliability of unstructured individual interviews was >37%.

Section 4 describes the steps taken in training and standardisation for raters of the speaking test to ensure rater reliability, and the analysis of rater reliability data. Sections 5 describes the steps taken to ensure test validity. These include test domain analysis and test content, the relationship between aviation training and the 'foundation' language used in the Checkpoint test, test taker characteristics and the relationship between language knowledge and subject matter knowledge, and the management of this in the Checkpoint test.

Section 6 describes two external research projects conducted by post-graduate students at Lancaster University, the key findings of which are:

- The 10-minute Checkpoint speaking test elicits >20% of the 3,000 most commonly occurring British and American English words as well as a range of broad aviation-related vocabulary with a significant difference in the number and type of words used by candidates at the different levels of proficiency;

- Scores on Checkpoint serve as valid predictors of the potential for linguistic success in aviation training. Put differently, students who score highly on Checkpoint should in general have little to no difficulty coping with the linguistic demands of English-medium initial aviation training.

This 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.

[1] Conway, J. M., Jako, R. A., & Goodman, D. F. (1995) *A meta-analysis of interrater and internal consistency reliability of selection interviews* Journal of Applied Psychology, 80, 565-579.

2. Introduction to reliability and validity

Validity

'Validity refers to the extent to which a test measures what it is intended to measure: it relates to the uses made of test scores and the ways in which test scores are interpreted, and is therefore always relative to test purpose'[2]. Thus, validity is understood to relate not only to a test itself, but also to the uses of the test and the inferences and decisions that are made on the basis of test scores.

Test validity in a particular testing domain is not absolute. It is easy to demonstrate invalidity, for example:

- A test of spoken Greek is not valid as a measure of English language reading skills; and
- A statistically unreliable test cannot be a valid measure.

However, it is not possible to prove validity. Rather, test developers make judgemental and empirical arguments for test validity following the principle that the more evidence there is for each category of validity, the stronger the overall validity argument will be. Validity categories include:

- Face validity – the judgement of users of the test, including aviation subject matter experts, that the test is appropriate to the domain and target skills.
- Construct validity – the extent to which performance on the test can be interpreted as a meaningful measure of language proficiency.
- Content validity – the extent to which test content is a representative sample of the domain-related language and skills.
- Predictive validity – the extent to which performance on a test is predictive of subsequent performance in non-test settings.

Decisions about the selection and use of a test should be made with consideration to the purposes for testing, a test's specifications and the strength of the argument for validity made by the test developers.

Reliability

Reliability is a measure of the ability of the test to minimise error so that scores are identical for candidates of identical ability, or equally, for the same candidates to receive the same scores if re-tested. Test reliability can be measured statistically.

Language tests set out to measure specific abilities, for example, listening skills or knowledge of vocabulary. We want variation in test scores to be linked to variation in test taker ability, and for the test to distribute candidates as far and as widely as possible with the lowest ability candidate receiving the lowest score and highest ability candidate receiving the highest score. However, factors which are not linked to language ability can affect test scores and are therefore sources of measurement error. These factors might be linked to the test itself such as test methods, differences in the different forms of the test or differences in rater behaviour. They may be linked to the test conditions, for example, administrative procedures or time of day. Or they may be linked to test taker characteristics unrelated to language proficiency such as age, first language and extent of subject matter knowledge. While it is accepted that some measurement error is inevitable, test developers seek to minimise measurement error in the design of tests so that variations in scores match variation in candidate ability as closely as possible.

Validity and reliability

Validity and reliability are often discussed as two separate, distinct qualities. In fact, they are inextricably linked: it is not possible for a test to be valid if the scores it produces are inconsistent. At the same time, reliability alone is

[2] Alderson, J.C., Clapham, C. & Wall, D. (1995) *Language test construction and evaluation* p.6 Cambridge: Cambridge University Press

insufficient to prove validity. For example, a test of vocabulary for elementary school children might achieve excellent reliability scores but would – obviously – be invalid for the purposes of selecting or admitting students for aviation training. Therefore, while reliability on its own is not enough, reliability is a necessary condition for test validity, and therefore, an evaluation of evidence for reliability should be the first consideration in the evaluation of a language test.

3. Checkpoint reliability

3.1 Internal test reliability

The most commonly used standard statistical measure for internal test reliability is the Cronbach alpha. Cronbach alpha scores are expressed on a scale from 0 to 1 where 1 indicates perfect internal reliability. If a test with perfect internal reliability were administered twice on the same candidates, it would produce the same distribution of scores and would rank the candidates in the same order so that correlation between the two sets of scores would be perfect. However, in the field of language testing there are many sources of potential measurement error and perfect internal reliability is rarely achieved[3]. Therefore, test developers work to established standards[4] for the acceptability of Cronbach alpha results:

Cronbach's alpha	Internal reliability
Above 0.90	Excellent
0.80 to 0.90	Good
0.70 to 0.80	Acceptable
0.60 to 0.70	Questionable
0.50 to 0.60	Poor
Below 0.50	Unacceptable

It should be noted that Cronbach's alpha is a lower bound estimate of reliability, so a result of 0.50 would mean test reliability somewhere between 50% and 100%.

To date, 702 candidates have taken the final versions of the A-set and B-set Checkpoint tests. Cronbach alpha results for these two final versions of the full Checkpoint test are:

- A set – **CA = 0.902** (test reliability between 90.2 and 100%) N=335
- B set – **CA = 0.915** (test reliability between 91.5 and 100%) N= 367

3.2 Standard error of measurement

The internal reliability estimates above tell us how reliable test scores are for a particular set of test takers. From these results a Standard Error of Measurement (SEM) can be calculated. The SEM tells us about the effect of measurement error and the extent to which we can have confidence in individual test scores. The SEM for green, yellow and red candidates in Checkpoint are reported in the tables below (please see the document '*Checkpoint - Test structure, platform and scores*' on the Latitude website for a description of the traffic light system). The figures show how a candidate's 'observed' score (in the test) may vary from the candidate's 'true' score (their actual language proficiency). For example, for candidates who score green in the A set, observed scores may vary from the candidate's true score by +/- 3.2%.

[3] Green, R (2013) *Statistical analysis for language testers* p.38 Basingstoke: Palgrave Macmillan

[4] George, D, & Mallery, P. (2003) *SPSS for windows step by step: A simple guide and reference*. p.231 (4th Edition) Boston: Allyn & Bacon

A set

Ability level	SEM
Green	+/- 3.2%
Yellow	+/- 1.8%
Red	+/- 2.8%

B set

Ability level	SEM
Green	+/- 3.0%
Yellow	+/- 1.7%
Red	+/- 2.5%

3.3 Correlation between the reading, listening and speaking tests

Correlation analysis provides information about the strength and direction of the relationship between two variables. Correlation is measures on a scale of -1 to 0 and 0 to +1. The table below shows correlations between scores in the Checkpoint listening, reading and speaking tests. The correlation coefficients show that there are significant (p<0.00, n=702) correlations between scores in the various parts of the test.

		Listening (%)	Reading (%)
Reading (%)	Pearson Correlation	**.668****	
	Sig. (2-tailed)	.000	
	N	702	
Speaking (overall)	Pearson Correlation	**.545****	**.553****
	Sig. (2-tailed)	.000	.000
	N	702	702

Although the results are highly statistically significant, the relationships between the reading, listening and speaking tests are moderate. As might be expected, the different skills are clearly related (they are all part of an overall English language ability). However, these results show that for individuals there may be a large variance between their ability in each area.

These results have implications for language testing in the aviation domain where it can be argued that knowledge of competence in all three skills is required to make valid selection and admissions decisions. The strong implication is that if you believe that proficiency in all three skills is required for successful aviation training, then all three skills must be individually measured in aviation selection tests. The alternative is already practiced in many selection interviews and oral tests where assessment is made on the basis of oral language proficiency alone. Even where a candidate is presented with listening or reading texts, if these require spoken answers then scores will be influenced by the candidate's speaking skills. In Checkpoint, speaking has a correlation of 0.55 (rounded) with both listening and reading. This means that only 55% of the variance in speaking scores can be attributed to either listening ability or reading ability. The implication from these results is that for any individual candidate, it will be impossible to accurately judge listening ability or reading ability from an oral test.

3.4 Reading and listening test item discrimination

Figure 1 below shows a typical result from our analysis of Checkpoint item (question) difficulty and discrimination. In figure 1, candidate scores for the listening test have been organised according to the traffic-light system (see the document '*Checkpoint – Test structure, platform and scores'* on the Latitude website).

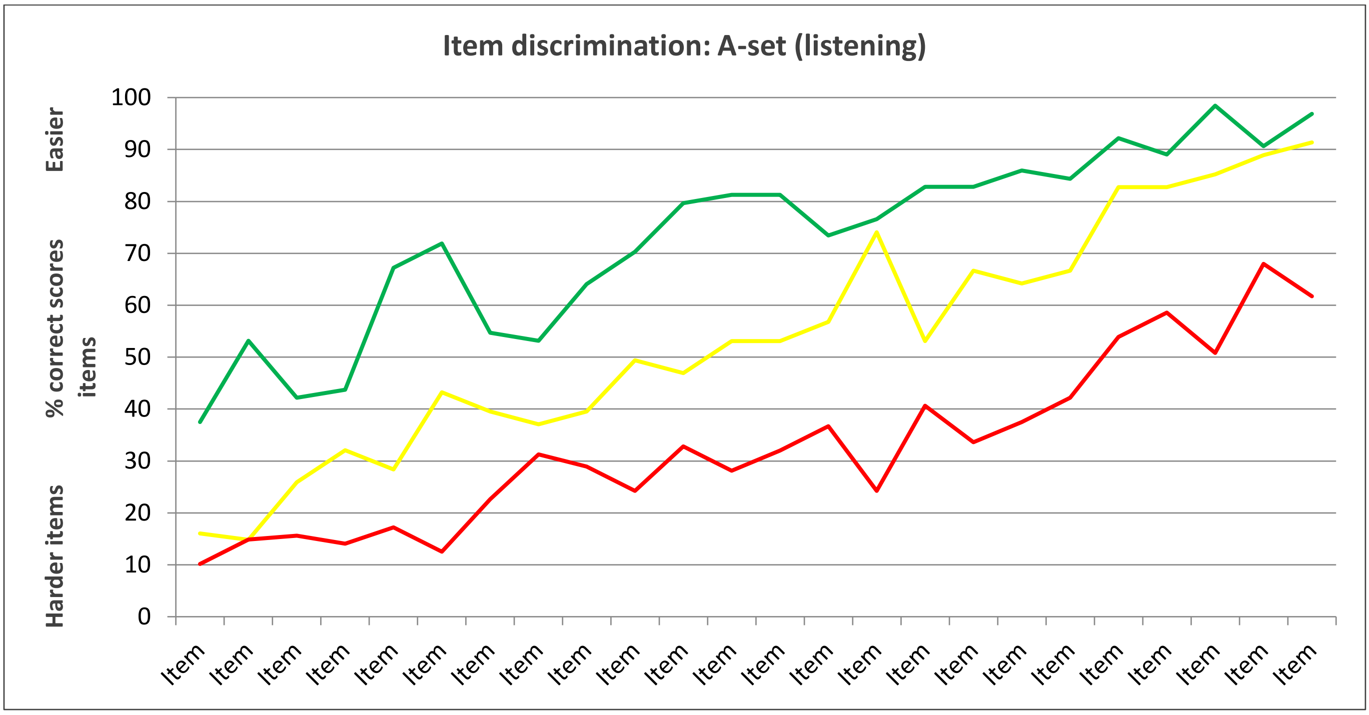


Figure 1: Item discrimination for the A –set Listening test

The item discrimination graph shows how good item (question) design, test trialling and item revision leads to a test with virtually perfect discrimination between the chosen candidate ability ranges. On the vertical axis is the test score expressed as a percentage. On the horizontal axis are 24 items belonging to a listening test. The aspects of good test design which we can see in the graph are:

- The test covers the complete range of item difficulty in a relatively smooth progression from more difficult items (on the left) to easier items (on the right). This is important for ability measurement: a test with too many easy items would not discriminate between high and medium ability candidates and, vice-versa, a test with too many difficult items would not discriminate between medium and low ability candidates.

- The more difficult items are on the left. Here, candidates of high ability (green) are scoring an average of 40% to 50% correct and both medium and low ability candidates are scoring below 20% correct. As expected, these more difficult items do not discriminate between medium and low ability candidates (the questions are too difficult for both) but these items are required in the test to clearly identify high and very high ability candidates. For example, 2 candidates might score "green", however, if one has an overall score of 95% and the other an overall score of 80%, it is clear which candidate has higher proficiency.

- The easier items are on the right. Here, candidates of low ability (red) are scoring an average of 50% to 60% correct and both medium and high ability candidates are scoring above 80% correct. As expected these easier items do not discriminate between medium and high ability candidates (the questions are too easy for both) but these items are required in the test to clearly identify low and very low ability candidates. For example, 2 candidates might score "red", however, if one has an overall score of 15% and the other an overall score of 40%, it is clear which candidate has higher proficiency.

- All the mid-range difficulty items discriminate extremely well between candidates in all three ability categories.

4. Rater training and standardisation

The Checkpoint speaking test is assessed by human raters according to an analytic rating scale. Checkpoint raters are aviation English language professionals, many of whom also have experience in aviation operations.

In order to generate reliable speaking test scores, all Checkpoint raters undergo thorough initial and recurrent rater training, and rater performance is monitored during field testing. Rater training takes place before each test administration following the steps outlined below. Rater training and standardisation material is accessible to raters at all times and is stored on secure password-protected pages on the Latitude website.

4.1 Rater training

Pre-standardisation

The rater:

- Familiarises his or herself with the test format by watching the test familiarisation videos and taking a Checkpoint test;
- Reviews the live speaking task battery;
- Reads the Checkpoint *Rater Guidance* document; and
- Listens to exemplar performances at red, yellow and green levels.

Standardisation

The rating team meets together with the rating team leader 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.

Certification

- Raters rate a set of five speech samples independently, returning scores to the team leader.
- The team leader analyses ratings and returns feedback on consistency and severity to the individual rater(s).

4.2 Rater reliability

Rater reliability is analysed using the Many Facet Rasch Analysis programme FACETS.[5] We accept only those raters who achieve infit mean square values of between 0.50 and 1.50 and, where applicable, outfit mean square values of between -2 and +2. Our current raters operate with in-fit values of between 0.94 and 1.22.

4.3 Monitoring raters

In-line with best practice, we analyse rater performance at periodic intervals during testing cycles. We 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. Any discrepancies are discussed with the rater in question.

[5] Linacre, 2016

5. Test validity

5.1 Test domain and test content

Language tests require context. Reading and listening comprehension tests require written and spoken 'texts' for candidates to process and respond to, and speaking tests need to present audio, textual and/or visual prompts in order to elicit a speech sample. In general language tests which aim to measure language in very broad language use domains, we find a wide range of input, for example, a bus timetable, a radio advertisement, newspaper classifieds or instructions for using a piece of garden machinery. In academic language tests used for decisions about university admission, the context is narrower and more specific. Candidates might be presented with a journal article on ornithology in Costa Rica, for example, or a lecture on the impact of government subsidies on agricultural practice. Whatever the context, the principle of validity is the same: 'we want to make inferences that generalise to those specific domains in which the test takers are likely to need to use language'[6]. In the domain of primary aviation training, scores need to be meaningful to decision makers and candidates, and to enable valid decisions about admission and preparatory language training to be made. Therefore, scores must reflect the ability of the student to cope with the real-life language use situations they will encounter when they arrive at the ATO. Based on actual observations of students and ATO staff engaged in training and training-related activities, and coupled with an analysis of syllabi and courseware for primary flight and ATC training, we recorded the key domain settings and language use tasks as follows:

Listening to:

- Instructors talking about technical subject matter in a formal classroom environment;
- Instructors talking with smaller groups, pairs or individual students about technical subject matter beyond the classroom, for example, in briefings and in simulator training facilities;
- Students in smaller groups or pairs talking about their training in less formal situations, such as in the cafeteria or rest area; and
- Students talking to training centre staff about issues related to training, for example, accommodation, visas and medical certificates.

Reading:

- Technical training textbooks and courseware;
- Aeronautical information manuals and publications;
- Regulatory documentation;
- Incident and accident reports; and
- Articles on aviation training, safety and management from industry websites, journals and periodicals.

Speaking about:

- Aeronautical mechanisms and processes;
- Events in aviation operations; and
- One's future career in aviation and about the aviation industry in general.

[6] Bachman, L. and Palmer, A. (1996) *Language Testing in Practice: Designing and Developing Useful Language Tests*, p44, Oxford: Oxford University Press

These settings and tasks are incorporated into the Checkpoint test with content carefully selected to represent the features and characteristics of the language as it is used in these settings. During test development, we engaged independent experts to collect judgements about the suitability of test content. This helped to ensure Checkpoint's content and context validity. This domain focus also improves face validity: it is considerably more engaging for candidates seeking a career in aviation to process and respond to content linked exclusively to aviation. This is not a trivial point: candidate engagement with the test materials improves focus and effort, and elicits better performances in tests which in turn improves response validity.

5.2 Test taker characteristics

General characteristics

Generally speaking, Checkpoint test takers:

- Are applicants, potential applicants or successful applicants for English-medium ab-initio flight or ATC training;
- Are young adults (typically between 18 and 28 years old);
- Are from a wide range of nationalities;
- Are sometimes highly motivated by the possibility of a career in aviation;
- Have received school-level education in Science, Technology, Engineering and Mathematics (STEM) subjects;
- May have received training in aviation or may have undergone under- and post-graduate education in STEM or aviation-related subjects, but are most likely to have little or no aviation subject matter knowledge.

Specific characteristics

Checkpoint test takers:

- Do not have English as a first language;
- Are from a wide range of first language backgrounds; and
- Vary widely in their level of ELP from beginner to fully competent user.

5.3 Aviation training and 'foundation' language

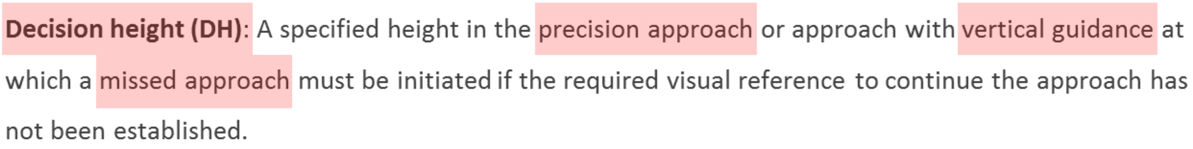
Professional aviation training is demanding. A typical initial training course for ATCOs begins with basic training comprising 12 weeks of classroom-led theoretical knowledge training. An EASA[7] integrated Airline Transport Pilot License course will begin with approximately 750 hours of classroom-led theoretical knowledge instruction and will require that students read 15 subject-specific textbooks with over 6,000 pages of training content. For students of aviation that have English as a first language (English L1), the task requires commitment and dedication. For those with the additional challenge of English as a second language (English L2), or even English as a third or fourth language, English Language Proficiency (ELP) can represent everything from an added layer of challenge to a full barrier to successful training. Of course, all students – both English L1 and English L2 - may struggle with training due to factors such as aptitude for learning, personal circumstances, volume of work and the student's study skills and ability to cope, but these issues are not related to ELP *per se*.

Generally speaking, students that have English as a first language (English L1) do not have language-related training issues if the training is delivered in English. Both English L1 students and their English L2 counterparts very often do not possess knowledge of aviation concepts or terminology. However, English L1 students are fully competent users of the

[7] European Aviation Safety Agency

language and have mastery of the vocabulary and grammatical structures common to formal written and spoken English as it appears in the teaching of Science, Technology, Engineering and Maths (STEM) subjects, what we at Latitude call 'foundation' language. This mastery of foundation language allows the acquisition of new aviation concepts and associated terminology simultaneously. This occurs as a natural, integrated part of the learning process as students listen to classroom instructors, read aviation textbooks and other learning materials, relying on foundation language to unlock meaning and to allow successful learning. As all students of aviation share many of the general characteristics above regardless of origin and L1, the obvious difference between English L1 and English L2 students is the level of ELP (hence the distinction between general and specific characteristics above).

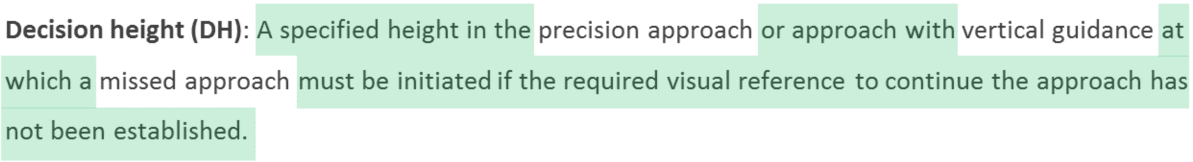
To illustrate this key difference between English L1 and English L2 students, below is a definition for the aviation-specific term 'decision height'[8]. Highlighted in red are those lexical items which have a specific meaning in the aviation domain:



Decision height (DH): A specified height in the precision approach or approach with vertical guidance at which a missed approach must be initiated if the required visual reference to continue the approach has not been established.

The terms highlighted in red and the concepts they represent will be new for *all* students regardless of first language. After all, learning these concepts is the reason why students follow ab-initio aviation training courses!

Here is the same definition of 'decision height', this time, with the 'foundation' language highlighted in green. This language is the formal, generic English common in the teaching of STEM subjects. This language is not exclusive to aviation, but is crucial to understanding aviation concepts and terminology such as 'decision height':



Decision height (DH): A specified height in the precision approach or approach with vertical guidance at which a missed approach must be initiated if the required visual reference to continue the approach has not been established.

This 'foundation' language is the language which is crucial for successful aviation training. English L1 students possess this language naturally and therefore ELP does not impede successful training. English L2 students may or may not have acquired this language before arriving at the ATO. Checkpoint is designed specifically to measure candidates' proficiency with foundation language and therefore provide a measure of the ability of English L2 students to cope with English-medium ab-initio aviation training.

5.4 Language knowledge and subject matter knowledge

The relationship between language knowledge and subject matter knowledge is a concern for language testers. Studies have shown that subject matter knowledge can make a difference in language test performance, with some test takers with subject matter knowledge having an advantage in language tests on one hand, and, on the other, those without subject matter knowledge being disadvantaged. The relationship between subject matter knowledge and language knowledge is complex: it appears to be linked to the level of difficulty and specificity of language tests and to the level of subject matter knowledge and language proficiency of test takers. In any case, as it is the test developers' desire to construct instruments that minimise measurement error, and therefore the issue of subject matter knowledge has been a central consideration in the development of Checkpoint for two key reasons:

1. As discussed above, in order to make valid inferences about the ability of students to cope once aviation training begins, it is essential to present test content and tasks which reflect the domain in question. Consequently, Checkpoint test content is geared exclusively to the domain of ab-initio aviation training.

[8] Source: ICAO Doc 4444, Procedures for Air Navigation Services, 1-6

2. Although the majority of Checkpoint candidates have little knowledge of aviation on entry to training, we recognise that some do (as specified in the general and specific test taker characteristics above).

Thus, there is a potential danger that varying degrees of candidate subject matter knowledge may lead to 'construct-irrelevant variance' in test scores. In order to control the effect of subject matter knowledge on test performance and minimise this variance, we have given great care to developing tasks and items which focus exclusively on candidates' knowledge of 'foundation' language as it appears in aviation training.

To do this:

1. We have expressly avoided developing tasks and items which require subject matter knowledge in order to respond correctly, i.e. the task cannot be successfully completed without subject matter knowledge. An example of the type if item we would reject for a speaking test might be '*Can you tell me how an altimeter works?*' or '*What are the advantages of synthetic flight training?*'
2. We have expressly avoided developing tasks and items which can elicit a correct response based on subject matter knowledge alone, i.e. without the need to process and respond to test content. An example of the type of item we would reject for a listening comprehension task might be presenting candidates with a recording of an instructor talking about pilot priorities when flying, along with an item requiring candidates to put the following words into the correct order:

   a. communicate b. aviate c. navigate

Any candidate with subject matter knowledge would be able to re-order these words correctly (b, c, a) without needing to process the recording. To illustrate how we control the effect of subject matter knowledge on test performance, we will look at some example tasks and items from the Checkpoint test.

5.5 Managing subject matter knowledge in the listening and reading tests

Below is an extract from a Checkpoint reading text[9] with the corresponding test task rubric and example item:

> Except for flights which are provided aerodrome control service only, the control of arriving and departing controlled flights shall be divided between units providing aerodrome control service and units providing approach control service …

Read the text and decide if the statements in the table are True (T), False (F) or Not Given (NG).

Tick (✔) the columns in the table.

	T	F	NG
1. Some flights do not receive both approach and aerodrome control services			

Regarding concern 1 above, the candidate does not need to know what an approach or aerodrome control service is in order to respond correctly. A correct response is perfectly possible based on reading ability and linguistic resource alone. The first sentence of the passage implies that some flights are provided with an aerodrome control service but not an approach control service, but a correct response – deciding if the statement is *true*, *false* or *not given* – requires processing of both the item statement and the complete passage, in particular, the first sentence and the meaning of the words 'except for' and 'only'.

[9] Source: ICAO Doc 4444, *Procedures for Air Traffic Management*, Section 4.3

Regarding concern 2 above, the statement in item 1 is a rule which is generally true. A candidate might possess this knowledge and therefore respond correctly. In reality, the statement is often false as many aerodromes provide both approach and aerodrome control services as standard (you cannot fly into most major airports without passing from approach to aerodrome control). Furthermore, even if the candidate possessed subject matter knowledge, he or she would need to process the passage in order to rule out the *not given* distractor by more than guesswork alone.

We have applied and carefully checked these two principles for managing subject matter knowledge in the development of each of the 48 items in each version of the Checkpoint listening and reading tests.

5.6 Managing subject matter knowledge in the speaking test

In task one of the Checkpoint speaking test, candidates describe an animation of an aeronautical mechanism or process. Candidates watch the animation twice, making notes if they wish, and on the third viewing, they describe the animation as they are watching. This is a complex task, the purpose of which is to elicit a sample of speech in order to assess:

- Fluency and competence with grammar, vocabulary and pronunciation;
- Functional competence, including the language of description, process, sequencing, linking of events, relationships between things, cause and effect etc.; and
- Strategic competence: The ability of the candidate to process visual representations of mechanisms and processes common in aviation with which they may or may not be familiar and to select the salient features of the animation and plan, organise and execute a response based on their linguistic resource.

Figure 1 shows a series of images that together represent the cycle of a 4-stroke engine, a typical task one animation in the Checkpoint speaking test. The cycle is clearly presented by the animation – no knowledge of combustion engines is required to respond to the task. Furthermore, the key technical vocabulary necessary to describe the animation, items which may be unknown to both English L1 and English L2 students, are clearly displayed on-screen, thus providing the lexical scaffold to perform the task successfully. The candidate's response, then, rests on the candidate's ability to describe the animation using his or her linguistic resource.

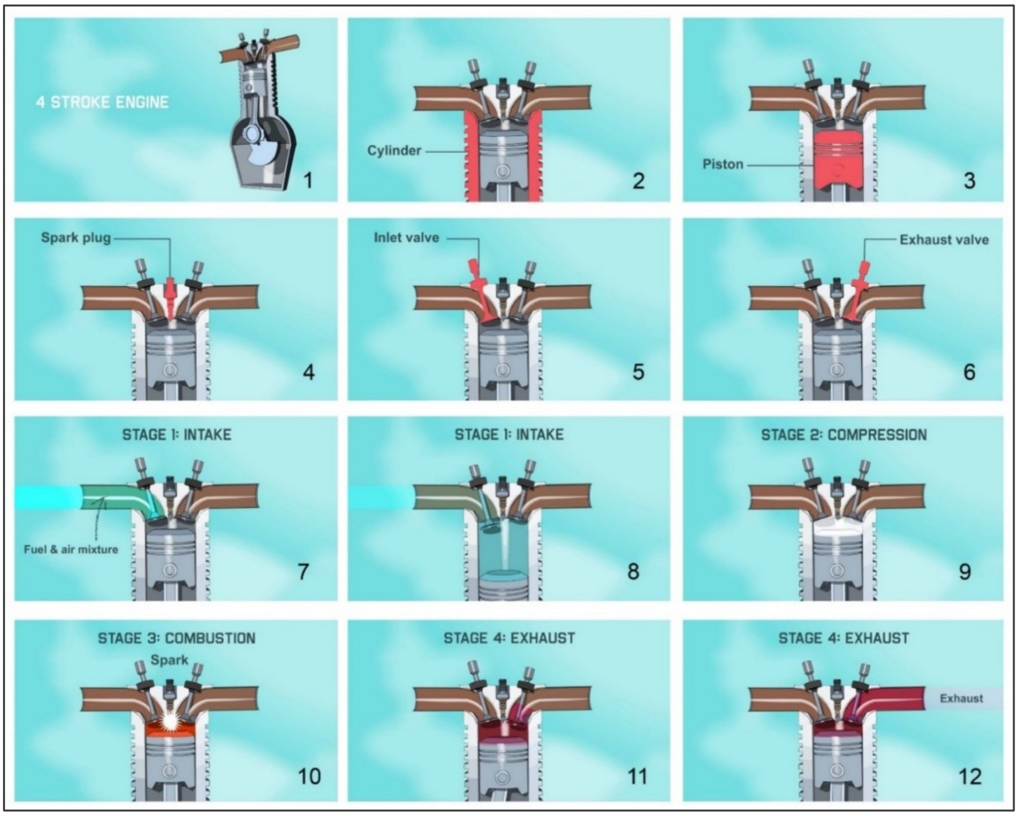


Figure 1: Checkpoint speaking task one: The cycle of a four-stroke engine.

5.7 Subject matter knowledge and Checkpoint scores

In both external review and in pre-testing and test trials, the feedback we received from aviation specialists was that subject matter knowledge would be essential in order to perform well in Checkpoint and that a lack of subject matter knowledge would not only impede test performance but may also demotivate students who wish to begin a career in aviation. To an extent, this was predictable in so far as subject specialists often perceive their area of expertise to be complex and therefore impenetrable to the layperson (indeed, language is a powerful agent in the creation of professional communities and notions of community 'insider' and 'outsider'). In any case, during pre-testing and test trials, in order to determine the extent to which subject matter knowledge affects Checkpoint test scores, we administered the test on test takers who were:

- Educated, adult English L1 and English L2 speakers including aviation and non-aviation specialists;
- English L1 undergraduate students at the Faculties of Business and Science and Engineering at the University of Plymouth; and
- English L2 students with a range of ELP levels attending pre-sessional and undergraduate courses at the Faculties of Business and Science and Engineering at the University of Plymouth.

The results showed that:

- Checkpoint consistently produced very high scores for English L1 non-aviation specialists. In several cases, English L1 non-specialists scored 100% in the listening and reading tests.
- Checkpoint produced a wide range of scores for English L2 non-specialists with a wide range of proficiency levels.
- Candidates with aviation subject matter knowledge were not advantaged in any way.

In subsequent field testing, the vast majority of whom did not have prior aviation training or subject matter knowledge, Checkpoint has produced scores which distribute candidates across a wide range of ability levels. Green[10] notes:

- The more homogenous the items are in terms of the construct being targets (language), the higher the level of internal reliability is likely to be; and
- Where items require a test taker to use his/her knowledge of maths, geography and so on in order to complete a language item, this may result in a weak level of internal reliability as the items will not be so closely related in terms of what is being targeted (language).

Analysis of the internal reliability of test scores from field testing shows that Checkpoint has high internal reliability estimates with homogenous test items (see section 3.1). Therefore, it is unlikely that any individual item is targeting construct irrelevant factors such as background knowledge. For this reason and those reasons outlined above, we are satisfied that the aviation specialists' perception that Checkpoint requires subject matter knowledge is not borne out by actual test performance.

[10] Green, R (2013) *Statistical analysis for language testers* p.39 Basingstoke: Palgrave Macmillan

6. Post-launch external investigations of test validity

We are committed to providing quality language testing services. We conduct research in order to understand how Checkpoint works, what scores mean, what impact the test has on our stakeholders, and how Checkpoint might be improved. In line with Latitude's research agenda, we are pleased make our test data available and to provide support to language testing students and researchers who wish to investigate aspects of Checkpoint's validity and reliability.

Since test launch in November 2014, two students on Lancaster University's Masters in Language Testing (Distance)[11] have conducted dissertation research on Checkpoint. Summaries of these two studies along with their key findings are summarised in the sections below.

6.1 Lexical output

Testing the speaking proficiency of ab-initio aviation industry trainees: an analysis of test takers' lexical output

Paul Sansom, 2015

Introduction

Deep analysis of oral test performances provides the opportunity to gather vital validity evidence and insight into how a test is performing. This study examined samples of test output in order to gather empirical evidence about the lexical characteristics elicited by the test at the red, yellow and green levels, and to compare candidate vocabulary with corpora of the most commonly occurring British and American English words. In addition, the study explored how the quantity and quality of lexical performance is affected by the three different task types.

Data and data analysis

30 test samples were selected to provide the data for this study including 11 green, 12 yellow and 7 red-rated performances for the vocabulary criterion. The test samples were transcribed and the transcriptions placed into the vocabulary profiler BNC-COCA.

A profile of lexical characteristics was generated from output taken from the entire sample population and presented using descriptive statistics. In addition, descriptive statistics of individual performance were calculated.

Transcriptions of green, yellow and red performances were separated to enable analysis of output at each proficiency level. Descriptive statistics were calculated and the means from each proficiency level are compared using one-way ANOVA and post-hoc calculations.

Transcriptions from each of the three tasks were analysed, allowing the impact of task characteristics on test takers' output to be evaluated with an examination of performance on each task at each level of proficiency.

Key findings

1. In the sample, Checkpoint elicited 626 of the 3,000 most commonly occurring British and American English words.
2. Checkpoint elicits a range of broad aviation-related vocabulary not found in the 3,000 most commonly occurring British and American English, for example, *cockpit* and *runway*.
3. There is evidence of a statistically significant difference in the number of words (tokens) families (headwords) and type (inflections, affixations) used by candidates at the different levels of proficiency.

[11] www.lancaster.ac.uk/linguistics/study/masters/courses/language-testing-distance-ma/#overview

4. There is wide statistical range and standard deviation within each proficiency level as well as large variation between candidates scoring green, yellow and red. The differences can primarily be accounted for by the variation in the number of tokens used at each level. Candidates scoring green use more tokens, resulting in more families and types being used. More differences occur between candidates scoring green and red, and to a similar extent between candidates scoring yellow and red. Differences between candidates scoring green and yellow were less marked.
5. Candidates scoring green:
    a. Use a marginally higher percentage of the 2,000 most commonly occurring English words than yellow or red candidates. This may support the view that proficient individuals possess the ability to use a solid base of high-frequency vocabulary;
    b. Use a greater number of 'sophisticated' words on average; and
    c. Use, on average, more prompt words (words presented in the task prompt) than test takers at other proficiency levels.
6. Taking fluency as a measure of task difficulty, task three appears to be the easiest task, eliciting a more fluent response (as measured by tokens per second), than tasks one and two. Task two appears to be the most difficult.

Recommendations

A more qualitative, discourse based approach may:

- Prove more revealing in understanding lexical performance at the three levels of proficiency than the purely qualitative approach used in this study; and
- Lead to the development of more empirically based rating scales that truly represent the type of lexical resource expected of test takers at each level of proficiency.

6.2 Predictive validity

A mixed-methods approach to assessing the predictive validity of a language proficiency test for aviation training admissions

Christopher Hamill, 2016

Introduction

Predictive validity can be defined as the degree to which some future (i.e. post-test administration) criterion measure can be predicted from a score on a prior assessment. Scores on a test said to be predictively valid for its stated use, therefore, should be approximately predictive of the value of the relevant criterion measure. As Checkpoint is a high-stakes entrance test, it is critical to investigate the predictive validity of using test scores for admissions in order to justify its continued use.

Some language tests report a composite score. Composite scores weight performance in all parts of a test in a single score and as such, may be seen as the fairest representation of overall candidate ELP. A further advantage of composite scores is that they are very easy to interpret and allow for streamlined decision making. In the development of Checkpoint, we conceived of listening, reading and speaking as language skills which are equally critical to success in aviation training. In other words, we believe that low candidate ELP in any of these skills would cause problems for students and their instructors and interfere with training efficiency and success. Therefore, a disadvantage of composite scores is that weak ELP in one skill may be masked by strong ELP in another, and therefore critical ELP issues may be obscured.

Checkpoint score reports detail candidate scores in each of the listening, reading and speaking tests along with advice on score interpretation and language training recommendations for those scoring red and yellow in any part of the test. We advise users to take into account *all* scores on each of the listening, reading and speaking tests when making selection and admissions decisions, and to be aware of the risks associated with weak ELP in any one skill. This helps to minimise Type I – or false positive – admissions decisions (i.e., those where students are admitted who should not have been). A possible disadvantage of this is that selection and admission decisions might be based on the lowest score and may fail to account for stronger ELP in other skills. Moreover, given that the overall speaking score is based on the lowest score in each of the five speaking criteria, decisions may be weighed most heavily on a single speaking criterion. For the purposes of the study, the lowest score in any part of the test was called the 'decision score'. The study, then, focussed on comparing the predictive validity of decision scores and composite scores, and to see if improvements could be made with Checkpoint score reporting and advice given to test users on interpretations of test scores.

Data and data analysis

Data[12] were collected from:

- 57 Checkpoint candidates who had been selected for and had begun integrated ATPL training; and
- 9 theoretical knowledge and flight instructors whom had taught or were currently teaching one or more of the students above.

The data were comprised of:

- Checkpoint test scores;
- Scores in internal and external aviation Theoretical Knowledge Examinations (TKEs); and
- Responses from student and instructor questionnaires and semi-structured interviews.

Quantitative data analyses included correlations between:

- Checkpoint scores (decision scores, composite scores and section scores) and scores in TKEs; and
- Checkpoint scores (decision scores, composite scores and section scores) and students' self-reported indicators of ELP for aviation training.

Topics for qualitative analyses included:

- The authenticity of Checkpoint;
- The importance of listening, reading, and speaking in aviation training; and
- ELP and TKEs.

Key findings

1. Scores on Checkpoint serve as valid predictors of the potential for linguistic success in aviation training. Put differently, students who score highly on Checkpoint should in general have little to no difficulty coping with the linguistic demands of English-medium initial aviation training.
2. Interviewees reported that the most important language skills for successful aviation training are reading, listening and speaking respectively.
3. Interviewees remarked unequivocally that Checkpoint's content was reflective of the domain-specific content and linguistic tasks performed during their subsequent aviation training. The students reported the impression

[12] Caveats:

- The data were restricted in range due to the fact that low-performing candidates were not admitted to flight training. Therefore, subsequent scores for aviation theoretical knowledge tests are only available for high-performing Checkpoint test takers. Such selection bias inevitably weakens correlations between Checkpoint scores and subsequent scores on Theoretical Knowledge Examinations (TKEs). If test takers who were ultimately not admitted to aviation training were somehow able to participate, one should expect to find more robust correlations between Checkpoint scores and these variables.
- The sample size was small. While data for all 57 students were available for correlations between scores on Checkpoint and TKEs, only 22 students' data were available for correlations between Checkpoint scores and the indicators of ELP for aviation training readiness. As a result, power for the statistical analyses involving these indicators was significantly reduced.

that the listening and speaking sections in particular corresponded with the linguistic demands of aviation training. One student even remarked that the content of a speaking item he encountered in Checkpoint was nearly identical to a conversation he later had with an instructor during the aviation training program.

4. Correlations between Checkpoint composite and decision scores and scores on TKEs were non-significant. This is unsurprising given that:
    a. Predictive validation studies frequently find only weak to moderate associations between language tests and measures representing post-test academic performance.
    b. Checkpoint is designed to measure ELP; TKEs are designed to measure aviation knowledge. Therefore, there is unlikely to be a strong association between scores on the two measures because they represent different constructs.
    c. Aviation training success requires proficiency in the language of instruction but involves considerably more than ELP alone. ELP is only one among many factors that affect academic success.

    NOTE: No internal reliability estimates of internal and external TKEs as measures of successful knowledge acquisition were available. Poor internal reliability of TKEs may have contributed to weak correlations between Checkpoint scores and TKE scores.
5. Significant correlations were found between listening and speaking section scores and students' self-reports of ELP readiness before and during aviation training.
6. Of the three skills measured in Checkpoint, correlations between reading scores and TKEs were the closest to significance. However, no significant correlations were found between Checkpoint reading scores and students' self-reports of ELP readiness before and during aviation training. The researcher reported that this may indicate that the Checkpoint reading test requires some improvement. However, we suggest that this may be explained, at least in part, by possible differences between the reading skills required to learn about aviation and the skills required to perform well in TKEs as measures of learning success. Although students reported that reading proficiency is extremely important given the volume of text that must be processed during training, it was also reported that advanced reading proficiency is not necessarily required for success in TKEs. Rather, what may be more important to success in TKEs is familiarity with the multiple-choice question format and the ability to cope with difficult and sometimes deliberately misleading questions. It is worth noting that students and their instructors reported that TKEs can be difficult even for English L1 students with full reading competence. More research into the reading skills required for success in TKEs is required.

Recommendations

To improve the usefulness of test scores, Latitude might consider:

1. Advising score users that considering the lowest score above all the others in selection and admissions decisions implicitly weights speaking performance most heavily; or
2. Modifying overall speaking test scoring to generate a numerical figure that incorporates all five criteria ratings rather than the existing overall traffic light score based on the lowest traffic light score in any criterion.

7. On-going investigations of test quality

As candidature grows, we will continue to engage with the language testing community and to invite postgraduate researchers to investigate Checkpoint according to the following broad research agenda:

- Establishing concurrence with other high stakes admissions tests;
- Further investigations into the relationship between language proficiency, subject matter knowledge and test performance;
- Investigating the validity of the listening and reading tests through:
  - Mapping reading and listening test items to language abilities;
  - Linking items to the CEFR based on expert judgement and test taker performance; and
  - Quantitative and qualitative analysis of the listening and reading texts;
- Investigating the validity of the cut scores for red, yellow and green levels;
- Investigating the validity of the speaking test through:
  - Qualitative analysis of candidate performance in part 3; and
  - Rater and test-taker introspection;
- Investigating bias; and
- Investigating the effect of Checkpoint on language learning (washback).