Revising the A2 Key and B1 Preliminary Listening exam

Alan Lanes
Occupational English Testing, Cambridge Assessment English

Brigita Séguis
Research and Thought Leadership, Cambridge Assessment English

Mark Elliott
Validation and Data Services, Cambridge Assessment English

CONTENTS

Introduction

Good practice in test development and validation requires regular review and evaluation of the exams to be conducted on an ongoing basis to ascertain that the underlying constructs of the test remain relevant and fit for purpose, and to identify areas where improvements are needed. In line with this approach, in 2014 Cambridge English embarked on a revision programme focusing on two of their lower-proficiency exams, A2 Key and B1 Preliminary (which were then known as *Cambridge English: Key* and *Cambridge English: Preliminary*, respectively), and their variants for schools. The focus of this article is the listening component of the two exams and the changes that have been made to A2 Key (hereafter, Key) and B1 Preliminary (hereafter, Preliminary) Listening papers and their variants for schools, A2 Key for Schools (hereafter, Key for Schools) and B1 Preliminary for Schools (hereafter, Preliminary for Schools).

Reviewing Listening test constructs

One of the main tasks carried out during the initial stage of test revisions was review of the constructs underpinning different parts of the test. Two developments, which took place since the exams were last revised in 2004, were taken into consideration during the review process, namely the introduction of upward certification and a move towards greater standardisation between Cambridge English Qualifications at different levels of proficiency to achieve greater continuity for learners and teachers.

The outcomes of the analysis of the Key and Preliminary Listening components should be interpreted with reference to the Cambridge English Cognitive Model for Listening Comprehension (Geranpayeh and Taylor (Eds) 2013), as well as the Common European Framework of Reference for Languages (CEFR, Council of Europe 2001) descriptors for listening comprehension at A2 and B1 levels; both of them are briefly outlined below.

The Cambridge English Cognitive Model for Listening Comprehension (Figure 1) perceives the listening process as comprised of five different levels of processing, namely:

- input decoding, when the listener transforms acoustic cues into groups of syllables
- lexical search, when the listener identifies the best word-level matches, based on a combination of perceptual information and word boundary cues
- syntactic parsing, when the lexical material is related to the co-text in which it occurs
- meaning construction, when general knowledge and inference are employed to add to the bare meaning of the message
- discourse construction, when the listener connects the new information to what was already known and decides how relevant it is.

The first three processes, namely input decoding, lexical search and syntactic parsing, are collectively described as 'lower-level processes' as they take place when a message is being encoded into language. The remaining two, namely meaning construction and discourse construction, can be classified as 'higher-level processes' since they are associated with building meaning.

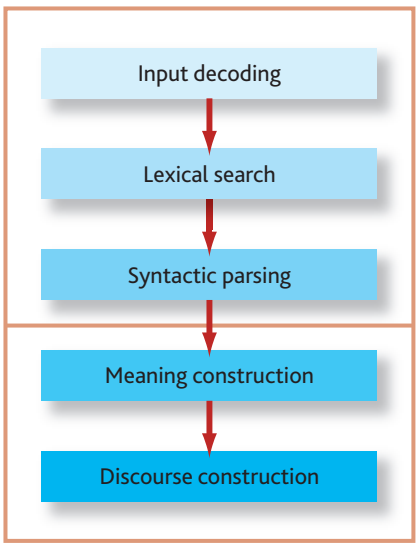

Figure 1: Cognitive processing model adapted from Geranpayeh and Taylor (Eds) 2013

The listening ability tested by Key and Preliminary exams spans across several levels on the CEFR, i.e. lower and higher A2 for Key, and lower and higher B1 for Preliminary. At the lower A2 level, the listener is expected to 'understand phrases and expressions related to areas of most immediate priority (e.g. very basic personal and family information, shopping, local geography, employment), provided speech is clearly and slowly articulated' (Council of Europe 2001:32). At the higher level of A2, the listener is expected to 'understand enough to be able to meet the needs of a concrete type provided speech is clearly and slowly articulated' (Council of Europe 2001:32).

As far as B1 level descriptors are concerned, the lower-level B1 descriptor states that a listener 'can understand the main points of clear standard speech on familiar matters regularly encountered in work, school, leisure etc., including short narratives' (Council of Europe 2001:66). At the higher B1 level, the listener 'can understand straightforward factual information about common everyday or job related topics, identifying both general messages and specific details, provided speech is clearly articulated in a generally familiar accent' (Council of Europe 2001:66).

The Cambridge English view is that test takers at A2 and B1 levels need to focus a great deal of attention at the more local levels of processing (input decoding, lexical search and syntactic parsing) and have little spare attentional capacity to give to the wider areas of meaning construction and discourse construction (Geranpayeh and Taylor (Eds) 2013). This had also been reflected in the design of the previous listening tasks of the Key and Preliminary tests, and had been revealed when the underlying task constructs were analysed. The analysis of the Key and Preliminary listening component also revealed one significant issue as far as construct coverage is concerned, namely the lack of test items that demand listening for gist.

During the analysis of the cognitive validity of the listening component of Cambridge English Qualifications, the extent to which different levels of cognitive processing are targeted in Key and Preliminary Listening sub-tests was investigated (Geranpayeh and Taylor (Eds) 2013). Following the analysis, it was concluded that, as far as lower-proficiency tests are concerned, there is a strong focus on perceptual-level processing. However, it should be borne in mind that the performance of lower-proficiency listeners, both in test conditions and in the real world, is largely conditioned by their ability to successfully employ compensatory strategies which enable them to infer general meaning even if the input has not been completely mastered. It would therefore seem relevant to include a number of items that would allow the test takers to demonstrate their ability to report the main point made by the speaker without grasping the full content of the message. In other words, what was missing from Key and Preliminary Listening sub-tests were items that demand listening for gist.

Following the publication of *Examining Listening* (Geranpayeh and Taylor (Eds) 2013), a number of changes to the Listening component of the Key and Preliminary tests have been implemented during the revision process. They are detailed in the two following sections of this paper.

Revised Key Listening test

A summary of changes made to the Key Listening test, including a comparison with the current version (until end of 2019) of the test, can be found in Table 1. The table charts the revisions through two trialling sessions.

Table 1: Revised Key Listening test format

Current: Key Listening (until end 2019)		*Trialing version 1: Key Listening*		*Trialing version 2: Key Listening Revised test format (from January 2020)*	
Timing	*Content*	*Timing*	*Content*	*Timing*	*Content*
22 minutes (approx.) plus 8 minutes transfer time	**Part 1** Five discrete 3-option multiple-choice items with visuals. 25–60 words.	29 minutes (approx.) plus 6 minutes transfer time	**Part 1** Six discrete 3-option multiple-choice items with visuals. Short neutral or informal dialogues. 40–60 words.	25 minutes (approx.) plus 6 minutes transfer time	**Part 1** Five discrete 3-option multiple-choice items with visuals. Short neutral or informal dialogues. 40–60 words.
	Part 2 Longer informal dialogue. Matching task. Five items and eight options. 150–170 words.		**Part 2** Longer informal dialogue. Matching task. Five items and eight options. 160–180 words.		**Part 2** Gap-fill. Longer neutral or informal monologue. Five gaps to fill with one word or a date or number or a time. 150–170 words.
	Part 3 Five 3-option multiple-choice items. 160–180 words.		**Part 3** Six 3-option multiple-choice items. Longer informal or neutral dialogue. 190–220 words.		**Part 3** Five 3-option multiple-choice items. Longer informal or neutral dialogue. 160–180 words.
	Part 4 Gap-fill. Five gaps to fill with one or more words or a number. 150–170 words.		**Part 4** Six discrete 3-option multiple-choice items with written options. Two or three B1 lexical/ structural items to be used to test candidate's understanding of the main idea, message, gist or topic. 40–60 words.		**Part 4** Five discrete 3-option multiple-choice items with written options. Two or three B1 lexical/ structural items to be used to test candidate's understanding of the main idea, message, gist or topic. 40–60 words.
	Part 5 Gap-fill. Five gaps to fill with one or more words or a number. 150–170 words.		**Part 5** Gap-fill. Longer neutral or informal monologue. Seven gaps to fill with one or two words or a number. 190–230 words.		**Part 5** Longer informal dialogue. Matching task. Five items and eight options. 160–180 words.

The most significant revision made to the test format was to Part 4 of the test, which was changed and trialled as discrete multiple-choice items that are aimed to test a candidate's understanding of the main idea, message, topic or gist, in line with the Cambridge English approach (Geranpayeh and Taylor (Eds) 2013). The addition of this task has allowed the construct of the Key Listening test to be expanded to include listening for gist.

In Phase 1 of the trialling, the number of items in the test was increased from 25 to 30 across the five parts of the test. This was to improve the accuracy and reliability of the test as well as to have a better coverage of the construct of listening comprehension at this level.

The main focus of this first trial was the newly designed Part 4 task – consisting of six discrete 3-option multiple-choice items with written options. There was also a change to the way the Part 3 task works. In the current test format, Part 3 is a cued dialogue which works on a need-to-know basis where one of the speakers cues in the questions and the other gives the key. This was amended so that the need-to-know basis was removed, yet questions are still cued in, but now by either speaker; both speakers now give the keys as well, thus better replicating a real-world dialogue between the two speakers. The range of question types was increased to test a candidate's ability to identify specific information, feelings and opinions.

Key and Key for Schools Listening trialling took place in the first quarter of 2016 in various locations and included the following language groups: Spanish, Portuguese, French, Polish, Russian, Serbian, Ukrainian, Dutch, Urdu and Malay. Several of the trials were carried out with post-assessment focus groups conducted by Assessment Managers linked to the various papers.

Results for the new Part 4 task

New Part 4 task results of the trial-tests can be seen in Table 2.

Table 2: Part 4 acceptance rates

Key	*Test 1*	*Test 2*	*Test 3*	*Test 4*	*Overall*
Number of trial-test candidates	140	135	168	194	637
Part 4 acceptance rate	100%	66%	83%	17%	66%

Key for Schools	*Test 1*	*Test 2*	*Test 3*	*Test 4*	*Overall*
Number of trial-test candidates	214	192	202	319	927
Part 4 acceptance rate	66%	83%	83%	0%	58%

All tasks were reviewed post-trial by a panel of experts who currently work on Key Listening, with decisions taken on whether the tasks were performing to the right level being based on both statistical information and expert judgement. Overall, 16 out of 24 (66%) for Key and 14 out of 24 (58%) for Key for Schools of the new Part 4 tasks were found to be at the right level and were accepted as suitable tasks for live test use. Tasks that were accepted at post-trial review were generally found to be limited in their reading load on the paper, limited on the complexity of the message itself, and on the complexity of language used within the script. Messages within the script were generally given in a linear manner with no backtracking from the speakers, and grammatical forms used were at the right level for candidates who can perform well at A2 level. Referencing within the texts needed to be clear and limited for the task to be accepted. The following accepted task is an example that is working at the required A2 level.

You hear two friends talking about going to university.

What subject is the man going to study?

A History
B Geography
C Chemistry

You hear two friends talking about going to university.

What subject is the man going to study?

F: So do you think you'll enjoy university?

M: Yes, especially the trips! I loved science at school, but I won't need it much. I thought about studying history but couldn't find a course I really liked. We're learning about mountains and rivers in the first term, which'll be great. And I'm looking forward to finding out about cities and how they've developed over time.

Those Part 4 tasks which were not accepted at post-trial review were found to have a very heavy processing cognitive load where either the amount of text on the question paper or the concepts in the scripts, or a combination of the two, proved too difficult for A2-level candidates. The following task was not accepted as it was found to be statistically too hard for candidates at A2 level.

You hear a woman talking to her boss, Jack, about a problem with a printer.

What's she done about the problem?

A She's ordered a new printer.
B She's asked her colleagues for advice.
C She's tried to repair the printer herself.

You hear a woman talking to her boss, Jack, about a problem with a printer.

What's she done about the problem?

F: Jack, can I talk to you about the printer?

M: Sure, is it still broken?

F: Yeah, I can see what's wrong with it, but I haven't managed to repair it.

M: Shall I just order a new one?

F: Maybe ... I could ask some colleagues for advice first if you like...

M: OK, that's a good idea.

F: Then I'll look at it again.

M: Thanks – great!

It became apparent during the review process that semantic matching tasks would need to have limitations on the reading load for the candidates, and this information was communicated back to the item writers.

There was discussion on whether the introductory sentences and questions should be recorded in the rubric or not, but the panel decided that rubrics should be read rather than leave a silence (as reading time would have to be allowed). Candidates may not be doing anything productive during this time. It was also found that the use of the present perfect and certain uses of modal verbs (e.g. should, might) may be too complex at times for the A2 candidates to process in connection with the timeframe.

There was also concern about the overall timing in the trial test and that if the entire rubric is read out for all parts of the test the timing will reach approximately 36/37 minutes. This may have given the impression that the test is now more difficult due to the fact that the time has increased by about 10 minutes on the current version of the test.

Key and Key for Schools questionnaire

527 questionnaire responses from candidates from 12 centres were captured and analysed. The majority of these respondents (77%) were preparing for Key and Key for Schools. Table 3 shows a breakdown of respondents by language.

Table 3: Main language groups in trialling

Language	*Number*	*%*
Spanish (including Catalan)	113	21
Russian	78	15
Polish	78	15
Chinese	51	10
Indonesian	35	7
French	33	6
Dutch	25	5
Bulgarian	22	4
Czech	20	4
Ukrainian	19	4
Portuguese	17	3
Other	36	7
Total	**527**	**100***

*does not sum to 100% due to rounding.

Candidate reactions to the tasks were generally positive, with the majority stating that the tasks were at the right level (Figure 2). Teachers were also positive about the tasks in the revised test (Figure 3).

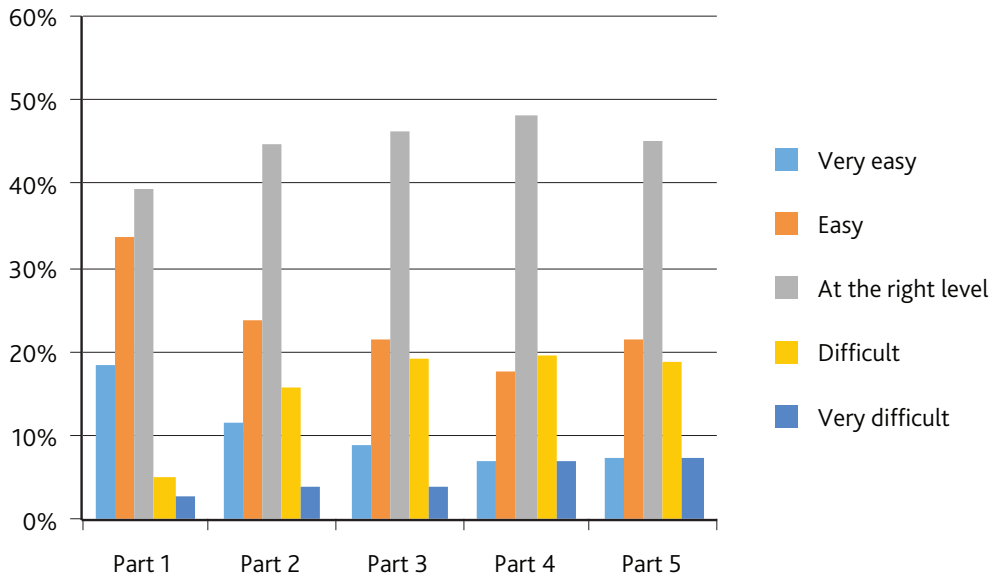

Figure 2: Candidates' perceptions of the difficulty of the test by part

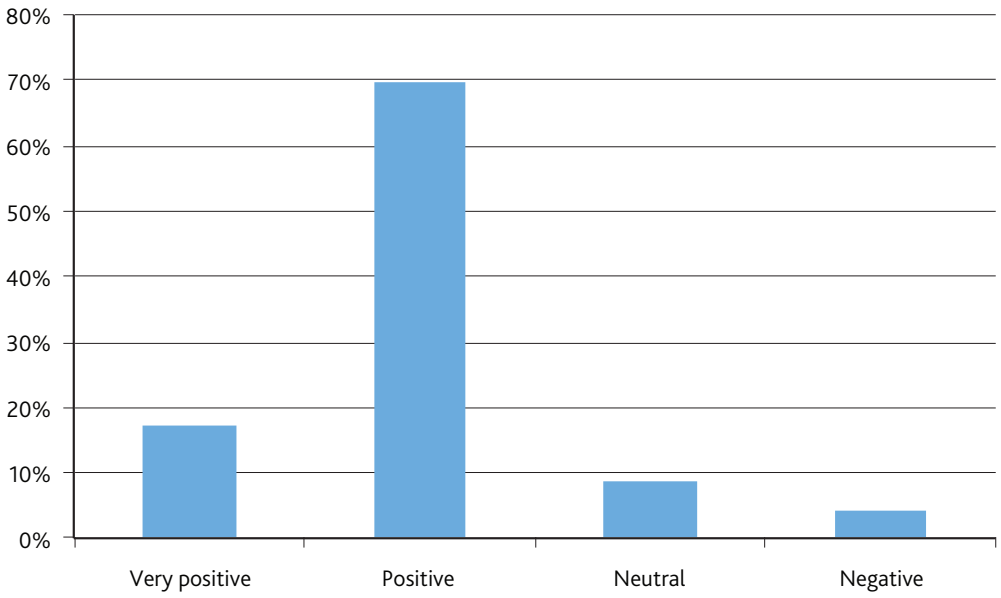

Figure 3: Teachers' perceptions of the revised test

Overall, 84% of candidates agreed that the test allowed them to show their listening ability. Between 78% and 100% of respondents reported that all tasks were appropriate for an A2-level test. Part 4 received the lowest rating of 78%; however, the vast majority of negative responses to this part were related to Version 4, whose Part 4 tasks were discovered to have a too high processing load and thus were not accepted as appropriate for the level.

Candidates and teachers were surveyed on what they liked and disliked about the tests. They commented that they particularly enjoyed the range of tasks and topics and that the new tasks may allow for better assessment of candidates' listening skills. One criticism was that the speed of delivery was too fast on some parts; this can be regulated during recording sessions. Another criticism was a lack of international accents. The revised test from 2020 will have a range of international English L1 accents.

Conclusions and recommendations from Key and Key for Schools Listening trial Phase 1

Based on the Phase 1 trialling results the following recommendations were proposed by the expert panel:

- Keep the new Part 4 tasks but ensure that sufficient item writer training is given so that the processing load and concepts are not above the level of the candidates.
- Reduce the number of items in the test (back to 25) as this would help to ensure that the time needed to take the Listening test does not significantly increase. As mentioned, an increase in time could wrongly lead the candidates to believe that the exam has become more difficult. The transfer time of six minutes should also remain as it is long enough to transfer 25 items to the answer sheet, only five of which are productive items.
- Use tasks in the second trial that are known to be working from Trial 1 in order to limit the need for a third trial; amendments should be made as appropriate to the longer Parts 3 and 5. This will mean tasks that have previously been calibrated can be used as anchor items across the trials.
- Change the order of the tasks to align Key and Key for Schools Listening with other Cambridge English Qualifications. The new order of the revised format is shown in Table 1, column 3, Trialling version 2.

Trial Phase 2

The second phase of the trial was set up using the new format of the test (Table 1, column 3, Trialling version 2) as recommended during the Phase 1 review. Items that were known to be performing at the level were chosen for the second phase of the trial. The aims of this second phase were to make sure that the test as a whole was working with the new test items and the updated 25-item format, and that candidates had enough reading time to read the questions, and enough time to transfer their answers in the six minutes now given at the end of the test.

Key and Key for Schools produced different versions of their tests. The Key for Schools version read out the full rubrics including all questions and scene setters, whereas the Key version did not read out the scene setters but gave the candidates silent reading time. Candidates' survey responses made no mention of whether the rubrics were read out or not and the items continued to show that they were performing as expected (in line with Phase 1). Answer keys were also tightened up so that only one word, or a number, or a date or a time (in line with the instructions on the revised test question paper) were allowed for the productive tasks (where candidates must write a response) to see what effect this might have on the statistics; but as there was only one trial test in this phase with one answer key the data from it was limited.

Revised Preliminary Listening test

As part of the wider revision of Preliminary and Preliminary for Schools, certain variants of the Listening components were trialled in order to explore possible changes to the current test design to better meet the needs of stakeholders.

Two formats of the new gist listening task were considered for inclusion in the test:

1. A 3-option multiple-choice format with text options.
2. A multiple-matching format with a shared pool of options for each item, similar to Part 3 of B2 First.

The testing focus of the two task versions is essentially the same, with the only significant difference being in the format. This difference, however, is a significant one within the context of a B1-level Listening test and the working memory demands placed on the listener – at B1 level, automatisation of the cognitive processes involved in listening is not well developed, meaning that listening places heavy demands on a learner's working memory (Geranpayeh and Taylor (Eds) 2013). A task format which places extra working memory demands on the listener as a result of the task format, rather than the underlying listening task, is likely to prove problematic and artificially difficult. Here, the second task format, which involves considering and selecting from a large number of written options while listening, is likely to create such extra working memory demands compared to a 3-option multiple-choice format where the task demands, although not trivial, involve reading or holding in working memory a much smaller set of options.

Eight test versions were trialled, featuring four versions of each task. The results of trialling supported the hypothesis on working memory demands; all four multiple-matching tasks returned calibrated Rasch difficulties which were well above the acceptable range for a Preliminary task, while all four 3-option multiple-choice tasks functioned well within acceptable limits (see Elliott and Stevenson 2015 for a discussion of how Cambridge English uses the Rasch model to create a sample-independent measurement scale of item difficulty for tasks).

The decision was therefore made to include the first task format (3-option multiple choice) in the revised Preliminary Listening test. This trial was one stage in a wider revision trialling process, as outlined in Figure 4.

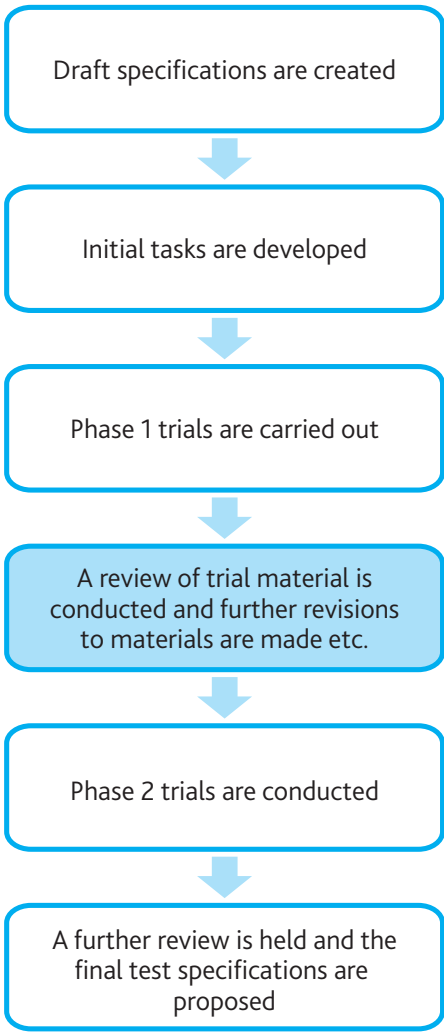

Figure 4: Revised test trialling process

Conclusion

The changes made to the Listening components of Key and Preliminary and their variants for schools were not as significant as those made to other components; however, the new formats assess a wider construct, namely introducing the assessment of listening for gist at this level. Changes were made to enable the tests to better align candidates in terms of upward certification and to offer greater standardisation across Cambridge English Qualifications.

References

Council of Europe (2001) *Common European Framework of References for Languages: Learning, Teaching, Assessment*, Cambridge: Cambridge University Press.

Elliott, M and Stevenson, L (2015) Grading and test equating, *Research Notes* 59, 14–20.

Geranpayeh, A and Taylor, L (Eds) *Examining Listening: Research and Practice in Assessing Second Language Listening*, Studies in Language Testing volume 35, Cambridge: UCLES/Cambridge University Press.