



Cambridge Assessment
English

Research Notes

Exam revisions 2020

Issue 76

November 2019

ISSN: 1756-509X



**Cambridge Assessment
English**

Research Notes

Issue 76/November 2019

Guest editors

Ardeshir Geranpayeh, Head of Automated Assessment & Learning,
Cambridge Assessment English

Ron Zeronis, Capability Owner Content Creation, Cambridge Assessment English

Editor

Sian Morgan, Senior Research Manager, Cambridge Assessment English

Production team

Anthony Power, Marketing Project Co-ordinator, Cambridge Assessment English

John Savage, Publications Assistant, Cambridge Assessment English

Typeset in the United Kingdom by George Hammond Design

Research Notes: 76

CONTENTS

2 Exam revisions as part of ongoing quality assurance processes

5 Consultation with English language teaching professionals to inform revisions to A2 Key and B1 Preliminary and their variants for schools

Debbie Howden and Angela Wright

12 Updating the A2 Key and B1 Preliminary vocabulary lists

Alan Lanes, Robbie Love, Bea Kalman, Mark Brenchley and Marianne Pickles

19 Revising the A2 Key and B1 Preliminary Listening exam

Alan Lanes, Brigita Séguis and Mark Elliott

29 Revising the A2 Key and B1 Preliminary Reading exam

Marianne Pickles, Tony Clark and Mark Elliott

39 Revising the A2 Key and B1 Preliminary Speaking exam

Kathryn Davies and Nick Glasson

56 Revising the A2 Key and B1 Preliminary Writing exam

Anthi Panagiotopoulou, James Lambie and Kevin Y F Cheung

Exam revisions as part of ongoing quality assurance processes

Revision of examinations is one of the stages of the test development model which has been practised in Cambridge Assessment English for decades. In this model, each exam is constantly monitored, reviewed and critically evaluated to identify areas for enhancement to ensure that they remain relevant to the needs of their stakeholders while taking advantage of the latest innovations in language learning, teaching and assessment. This process may prompt revisions to the exams with the desire to provide better feedback for learners and other stakeholders. In issue 62 of *Research Notes* we reported on the revisions to *Cambridge English: First* and *Cambridge English: Advanced* (now known as B2 First and C1 Advanced). In this issue we report on the revisions to A2 Key and B1 Preliminary (henceforth Key and Preliminary) and their variants for schools.

In the first article, Howden and Wright set the scene for change by reporting how Cambridge Assessment English surveyed various stakeholders about their satisfaction with Key and Preliminary and their variants for schools. A range of English language teaching professionals including Centre Exams Managers, Directors of Studies, teachers and Examination Administrators were surveyed to find out their overall satisfaction levels with the exams under review. Although the overall satisfaction was very high and did not in itself suggest significant need for revisions to the exams, it highlighted areas for improvement. Following the survey, a revision project team was set up to look at possible improvements to Key and Preliminary and their variants for schools. The team focused on the four skills of listening, speaking, writing and reading, in addition to reviewing the wordlists used in all skills.

Cambridge English has, in recent years, been increasingly taking advantage of corpora-based methodologies to create wordlists for its exams, which are based on the levels of the Common European Framework of Reference for Languages (CEFR). Lanes,

Love, Kalman, Brenchley and Pickles explain how Key and Preliminary wordlists were updated to provide better validity evidence of the appropriate language use for these exams. They explain how the revised A2 and B1 lists expanded the breadth of lexis that would be available across a range of functions, genres and topics mandated by the revised specifications, and discuss a three-stage revision procedure for making this update. The revised vocabulary wordlists for Key and Preliminary are aimed at continuous support for the evolving learning and assessment needs of learners at these levels.

Lanes, Séguis and Elliott describe the process of changing the Listening tests to reflect the latest theories outlined in *Examining Listening: Research and Practice in Assessing Second Language Listening* (Geranpayeh and Taylor (Eds) 2013). They report on examining the constructs underpinning different parts of the Listening tests and changing their sub-constructs to not only cover a wider construct but also to further align them with other Cambridge English Qualifications, both for upward certification and greater standardisation across the whole exam suite.

One of the considerations in the revisions of Key and Preliminary was reporting Reading and Writing scores independently. There has been a long-running debate about what constitutes specifically writing skills at A2 and B1 levels in light of the overlap of the underlying test constructs for tasks designed to measure Reading and Writing at these levels. Key and Preliminary were no exception and had combined reading and writing tasks in one paper since their inception. There was a request from some of our stakeholders to report separate scores for these skills. To achieve this, we had to make changes to some of the tasks in Key and Preliminary. This resulted in separate papers of Reading and Writing for Preliminary. Key, despite new task designs and separate reporting of scores, still combines Reading and Writing in the same paper for practical reasons. Pickles, Clark and Elliott explain

these changes, arguing that introducing new tasks in Key and Preliminary, particularly the introduction of expeditious reading in Key, better reflects current understanding of the reading construct.

Perhaps the biggest change in the revised exams was introduced in the Writing sections. Writing in the Key exams was assessed as part of the Reading and Writing papers and the performance for these two skills was reported together. In the revised Key Writing, some of the tasks, such as spelling, were removed, and others, such as open cloze, moved to the Reading section. The guided writing task remained but has been modified to cover performance up to B1 level, and we now have a new writing task, picture story, that allows us to assess A2 writing for narrative rather than simple transactional purposes. Panagiotopoulou, Lambie and Cheung describe these changes, and also explain how the Preliminary Writing section is now an independent paper with revised writing tasks that allows us to measure various aspects of writing at this level. There is now a separate paper in Preliminary for Writing comprising new tasks, which will bring much more alignment for score reporting of Cambridge English Qualifications across all CEFR levels from A2 to C2.

The changes to the Speaking components of Key and Preliminary are described by Davies and Glasson. They explain how the new format for Key Speaking elicits a wider range of language and language functions to allow candidates to fully demonstrate their speaking skills, and provides a more authentic and meaningful task. They highlight the production of new additional supplementary materials to prepare Speaking Examiners to provide better scaffolding for the learners which will in turn have the positive washback effect of learners demonstrating their speaking skills fully. For Preliminary Speaking, they explain that the new test provides much greater interlocutor control than the previous test design. The revised Preliminary Speaking test allows stronger candidates to show a fuller range of skills and at the same time aims to support less able candidates more than previously.

Overall, the revisions of A2 Key and B1 Preliminary and their variants for schools ensure that the test constructs remain fit for purpose, especially in relation to upward certification and standardisation with Cambridge English Qualifications at B2, C1 and C2.

Consultation with English language teaching professionals to inform revisions to A2 Key and B1 Preliminary and their variants for schools

Debbie Howden Business and Marketing, Cambridge Assessment English

Angela Wright Business and Marketing, Cambridge Assessment English

Introduction

At Cambridge Assessment English we review our exams on a regular basis to ensure that they remain relevant to the needs of learners and schools and incorporate evolving approaches to assessment and learning. The first step in the revision process is to consult English language teaching professionals globally, who have experience of running the exams or preparing learners for them. In this article, we report on the stakeholders' consultation survey we carried out regarding the revisions to A2 Key and B1 Preliminary and their variants for schools, which were then known as *Cambridge English: Key* and *Cambridge English: Preliminary*, respectively.

We surveyed a range of professionals in order to understand their satisfaction with the exams and their recommendations for improvements. The survey included some open-ended questions to give respondents the opportunity to explain the reasons for their responses. The survey found that overall satisfaction levels with the exams were very high and did not indicate that significant revisions were required. However, the feedback received from this study helped to inform a wider review of the exams by assessment experts from within Cambridge Assessment English and consultants.

Methodology and respondent profile

907 English language teaching professionals around the world, including Centre Exams Managers, Directors of Studies, Heads of English, teachers and Examination Administrators, participated in an online survey. Over half of the responses (55%) were from English language teachers. 596 respondents were currently running or preparing learners for A2 Key/A2 Key for Schools and 761 were currently running or preparing learners for B1 Preliminary/B1 Preliminary for Schools at their institutions.

Findings

The survey aimed to find out what stakeholders like about the existing exams and what works well, and to identify potential areas for improvement, particularly with regard to:

- test content
- test length
- task types
- results reporting
- preparation material – the amount of preparation material available and the quality.

Please note that where the total percentages shown in the tables below do not add up to 100, this is due to rounding.

Test content

Respondents were asked how satisfied (on a 5-point scale from very satisfied to very dissatisfied) they were with the appropriacy of topics and the variety of topics, to ensure the exams remain fit for purpose for the target audiences of school-age and adult learners. Overall, satisfaction was very high, with 86%+ either quite satisfied or very satisfied with the appropriacy of topics (see Table 1) and 83%+ either quite satisfied or very satisfied with the variety of topics across each of the exams (see Table 2). Typical comments in relation to the test content were (note comments are unedited to maintain authenticity):

'Topics are suitable for the students' age.' (Teacher, Italy).

'I like that the topics have to do with the students' daily life, and that it tests the four skills.'
(Teacher, Switzerland)

Table 1: Appropriacy of topics

	<i>Very satisfied</i>	<i>Quite satisfied</i>	<i>Neither satisfied nor dissatisfied</i>	<i>Quite dissatisfied</i>	<i>Very dissatisfied</i>
A2 Key	29%	61%	8%	2%	0%
A2 Key for Schools	33%	53%	12%	1%	0%
B1 Preliminary	27%	61%	10%	2%	0%
B1 Preliminary for Schools	31%	56%	12%	1%	0%

Base: A2 Key (n=231), A2 Key for Schools (n=374), B1 Preliminary (n=381), B1 Preliminary for Schools (n=471)

Table 2: Variety of topics

	<i>Very satisfied</i>	<i>Quite satisfied</i>	<i>Neither satisfied nor dissatisfied</i>	<i>Quite dissatisfied</i>	<i>Very dissatisfied</i>
A2 Key	30%	57%	11%	2%	0%
A2 Key for Schools	31%	54%	13%	1%	0%
B1 Preliminary	31%	54%	13%	2%	0%
B1 Preliminary for Schools	30%	53%	16%	1%	0%

Base: A2 Key (n=229), A2 Key for Schools (n=376), B1 Preliminary (n=378), B1 Preliminary for Schools (n=471)

All Cambridge English Qualifications assess all four skills (reading, writing, listening and speaking). When asked how important it was to assess all four skills, respondents were overwhelmingly in favour of continuing to assess all four skills, with 91% of respondents indicating that this is very important (see Table 3).

Table 3: Importance of assessing all four skills

	<i>Very important</i>	<i>Quite important</i>	<i>Neither important nor unimportant</i>	<i>Not important</i>	<i>Not important at all</i>
All four skills	91%	6%	2%	0%	1%

Base: All respondents (n=840)

Length of the tests

It is important to find a balance between assessing all skills adequately and a length that is appropriate for the level and age of candidates. To establish respondents' satisfaction with the length of the test, we asked them two questions for each of the papers: (1) Is the number of questions too many, just right or too few? (2) Is the time given to complete the exam too long, just right or too short?

While satisfaction levels with the length of the Listening and Speaking papers were very high (85%+), it was felt that there was some room for improvement in the Reading and Writing papers, particularly for B1 Preliminary/ B1 Preliminary for Schools. This has been taken into account with the revised exams. In B1 Preliminary/ B1 Preliminary for Schools, Reading and Writing have been split into separate papers. For A2 Key/A2 Key for Schools, Reading and Writing remains as a combined paper but the format has changed from 56 questions in nine parts to 30 Reading questions in five parts and two Writing parts (see Tables 4–11). There was no significant difference in responses between the exams for schools and the standard exams.

Table 4: Number of questions for A2 Key

<i>A2 Key</i>	<i>Reading and Writing paper</i>	<i>Speaking paper</i>	<i>Listening paper</i>
Too many	14%	1%	5%
Just right	83%	85%	93%
Too few	2%	14%	2%

Base: All respondents (n=303–312)

Table 5: Number of questions for A2 Key for Schools

<i>A2 Key for Schools</i>	<i>Reading and Writing paper</i>	<i>Speaking paper</i>	<i>Listening paper</i>
Too many	18%	2%	8%
Just right	80%	87%	90%
Too few	2%	12%	2%

Base: All respondents (n=447–464)

Table 6: Number of questions for B1 Preliminary

<i>B1 Preliminary</i>	<i>Reading and Writing paper</i>	<i>Speaking paper</i>	<i>Listening paper</i>
Too many	18%	2%	9%
Just right	78%	93%	89%
Too few	3%	5%	2%

Base: All respondents (n=464–474)

Table 7: Number of questions for B1 Preliminary for Schools

<i>B1 Preliminary for Schools</i>	<i>Reading and Writing paper</i>	<i>Speaking paper</i>	<i>Listening paper</i>
Too many	23%	3%	11%
Just right	75%	91%	87%
Too few	3%	5%	2%

Base: All respondents (n=546–555)

Table 8: Exam length – A2 Key

<i>A2 Key</i>	<i>Reading and Writing paper</i>	<i>Speaking paper</i>	<i>Listening paper</i>
Too long	5%	2%	2%
Just right	83%	92%	94%
Too short	12%	6%	4%

Base: All respondents (n= 303–310)

Table 9: Exam length – A2 Key for Schools

<i>A2 Key for Schools</i>	<i>Reading and Writing paper</i>	<i>Speaking paper</i>	<i>Listening paper</i>
Too long	8%	2%	4%
Just right	83%	92%	91%
Too short	10%	6%	5%

Base: All respondents (n=455–460)

Table 10: Exam length – B1 Preliminary

<i>B1 Preliminary</i>	<i>Reading and Writing paper</i>	<i>Speaking paper</i>	<i>Listening paper</i>
Too long	5%	1%	3%
Just right	74%	93%	92%
Too short	21%	6%	5%

Base: All respondents (n=460–473)

Table 11: Exam length – B1 Preliminary for Schools

<i>B1 Preliminary for Schools</i>	<i>Reading and Writing paper</i>	<i>Speaking paper</i>	<i>Listening paper</i>
Too long	7%	2%	5%
Just right	75%	94%	89%
Too short	18%	5%	7%

Base: All respondents (n=533–548)

Task types

Respondents were asked to rate their satisfaction with each task type in each of the papers to identify if there were any that users felt didn't work. Respondents indicated high levels of satisfaction across all of the papers, with at least 70% reporting they were satisfied or very satisfied with almost all task types. Some of the reading and writing tasks in A2 Key/A2 Key for Schools had slightly lower satisfaction levels, and two of the tasks with the lowest satisfaction ratings have been dropped from the revised exams: Part 6 – Word completion (66% satisfied or very satisfied for A2 Key, 68% for A2 Key for Schools) and Part 8 – Information transfer (66% satisfied or very satisfied for A2 Key, 69% for A2 Key for Schools). This feedback supported expert opinion in deciding which task types to include in the revised exams. The revisions also aimed to ensure more overlap of task types between A2 Key/A2 Key for Schools, B1 Preliminary/B1 Preliminary for Schools and higher-level Cambridge English Qualifications, where appropriate, to give learners who progress through the exams a greater sense of development.

Score reporting

The survey also asked about overall satisfaction with results reporting and about which aspects of reporting are most important for users.

Most respondents were satisfied with how results are reported across each of the exams (67%+). Respondents indicated that it is important to provide not only an overall result (94%), but also scores by skill (91%), Common European Framework of Reference for Languages (CEFR) level (89%) and an indication of which skills require improvement (85%). A majority also felt it important to keep reporting results at the CEFR level above (77%) and below (70%). We decided, therefore, to keep results reporting as it is now (see Tables 12 and 13).

However, based on a separate survey with a similar audience of nearly 1,000 teachers, Heads of English and Centre Exams Managers, which looked at how grades are reported, it was decided to change from Pass with Distinction/Pass with Merit/Pass to grades A/B/C, in line with higher levels of Cambridge English Qualifications. A typical response was:

'Since students tend to do not only KET [A2 Key] or PET [B1 Preliminary] exams but normally work towards FCE [B2 First] or CAE [C1 Advanced] you should have the same grading scales for all exams.'
(Head of English, Portugal)

Table 12: Satisfaction with how results are reported

	<i>Very satisfied</i>	<i>Quite satisfied</i>	<i>Neither satisfied nor dissatisfied</i>	<i>Quite dissatisfied</i>	<i>Very dissatisfied</i>
A2 Key	29%	40%	15%	9%	6%
A2 Key for Schools	28%	39%	17%	11%	5%
B1 Preliminary	27%	41%	15%	11%	5%
B1 Preliminary for Schools	27%	42%	16%	11%	5%

Base: A2 Key (n=319), A2 Key for Schools (n=474), B1 Preliminary (n=484), B1 Preliminary for Schools (n=566)

Table 13: Importance of reporting

<i>Across all exams</i>	<i>Very important</i>	<i>Quite important</i>	<i>Neither important nor unimportant</i>	<i>Not important</i>	<i>Not important at all</i>
Overall result on all skills	83%	11%	5%	1%	0%
Scores broken down by skill	77%	14%	7%	1%	1%
CEFR level	73%	16%	9%	2%	1%
An indication of which skills require improvement	70%	15%	11%	2%	2%
Reporting results above CEFR level	57%	20%	16%	4%	3%
Reporting results below CEFR level	47%	23%	20%	5%	5%

Base: All respondents (n=537–575)

Preparation

Respondents were fairly satisfied (59%+ satisfied or very satisfied) with the amount of preparation material available (see Table 14). They were also fairly satisfied (56%+ satisfied or very satisfied) with the quality of free preparation materials available (see Table 15). However, it was clear that there was room to improve both the availability and quality of the free materials to help learners prepare for the exams. A separate large-scale survey of preparation centres for Cambridge English Qualifications has suggested that the issue is lack of awareness of the support materials available rather than not enough support. Work is, therefore, being undertaken to make the support more easily accessible and to keep schools and teachers informed about new materials. All exam support materials are being updated to reflect the exam revisions and to ensure that they are of appropriate quality.

Table 14: Amount of preparation materials available

	<i>Very satisfied</i>	<i>Quite satisfied</i>	<i>Neither satisfied nor dissatisfied</i>	<i>Quite dissatisfied</i>	<i>Very dissatisfied</i>
A2 Key	24%	35%	25%	15%	2%
A2 Key for Schools	22%	45%	21%	10%	2%
B1 Preliminary	25%	41%	23%	9%	1%
B1 Preliminary for Schools	22%	46%	22%	9%	1%

Base: A2 Key (n=232), A2 Key for Schools (n=379), B1 Preliminary (n=377), B1 Preliminary for Schools (n=471)

Table 15: Quality of free preparation material

	<i>Very satisfied</i>	<i>Quite satisfied</i>	<i>Neither satisfied nor dissatisfied</i>	<i>Quite dissatisfied</i>	<i>Very dissatisfied</i>
A2 Key	18%	38%	28%	11%	4%
A2 Key for Schools	22%	36%	27%	12%	4%
B1 Preliminary	23%	39%	26%	10%	2%
B1 Preliminary for Schools	22%	40%	26%	10%	2%

Base: A2 Key (n=223), A2 Key for Schools (n=365), B1 Preliminary (n=367), B1 Preliminary for Schools (n=471)

Conclusions

Overall, there was very high satisfaction with A2 Key/A2 Key for Schools and B1 Preliminary/B1 Preliminary for Schools among the English teaching professionals who use the exams. Most of the changes made to the exams for the 2020 revisions are, therefore, based on expert opinion and latest thinking in testing methodology rather than user feedback to ensure they remain relevant.

In addition to the statistical data, respondents commented on reasons why they like the exams. For example:

'I like Key because it provides an international standard; students like the experience of doing the test and using real, every day English.' (Centre Exams Manager, Brazil)

'The exam gives the students a good first-hand experience at exam taking and allows their growth in self-confidence towards higher levels.' (Teacher, Argentina)

'It's what students and parents choose. Not only do they want to learn English but it is important to gain a qualification that will help them with further studies.' (Centre Exams Manager, Italy)

'It is a reliable exam that allows institutions and universities have a clear idea of their students' English level.' (Centre Exams Manager, Colombia)

'Key for Schools makes candidates feel ready to take on the world.' (School English Coordinator, Bolivia)

The main area where it was felt that improvements could be made was not in the exams themselves but with the support material. This is being addressed through a review of all the exam support materials, and ensuring that the full range of support offered is promoted more effectively and made more easily accessible.

Updating the A2 Key and B1 Preliminary vocabulary lists

Alan Lanes Occupational English Testing, Cambridge Assessment English

Robbie Love School of Education, University of Leeds

Bea Kalman ELT Technology, Cambridge University Press

Mark Brenchley Research and Thought Leadership, Cambridge Assessment English

Marianne Pickles New Product Development, Cambridge Assessment English

Introduction

Since 2006, Cambridge Assessment English has published vocabulary lists for four exams from its Cambridge English Qualifications: A2 Key, A2 Key for Schools, B1 Preliminary and B1 Preliminary for Schools. This article describes the process of updating these lists as part of the revised 2020 specifications for all four exams.

Background to the lists

Cambridge English currently publishes two vocabulary lists for the four Cambridge English Qualifications that target the lower end of the Common European Framework of Reference for Languages (CEFR, Council of Europe 2001, 2018) scale: the Key Vocabulary List, which covers both A2 Key and A2 Key for Schools, and the Preliminary Vocabulary List, which covers both B1 Preliminary and B1 Preliminary for Schools.¹

Both lists serve parallel functions. Firstly, they provide item writers with an explicit set of vocabulary that is to form the core of all A2 Key (hereafter, Key) and B1 Preliminary (hereafter, Preliminary) tasks. Secondly, they enable students and teachers to identify the core vocabulary that they should target to help ensure success in the exam. Both functions reflect the more fundamental goal of ensuring that the exams accurately and fairly assess learners at the targeted CEFR levels: A2 in the case of Key and B1 in the case of Preliminary. Learners at these lower CEFR levels tend to have access to a restricted lexical repertoire, and find it harder to go beyond this repertoire to demonstrate higher-level language behaviour, such as the capacity to infer meaning from context (Docherty 2015). Accordingly, the provision of vocabulary lists helps ensure not only that the Key and Preliminary exams draw on an appropriate lexical repertoire for the targeted CEFR level, but that teachers and students have access to the repertoire that candidates can reasonably be expected to draw on.

Historically, the content of the original Key and Preliminary lists was based on the intuitions of language experts, with the Key list primarily drawing on the vocabulary detailed in the Council of Europe's *Waystage 1990* specifications (Van Ek and Trim 1991b) and the Preliminary list drawing on the vocabulary detailed in the *Threshold 1990* specifications (Van Ek and Trim 1991a). As part of the ongoing review process, later versions of the vocabulary lists have increasingly drawn on more objective sources of information; in particular, corpora, which are carefully curated, representative collections of language use (Ball 2002, Street and Ingham 2007). Corpora constitute an invaluable evidence base, enabling assessment specialists to more accurately identify what language use actually looks like when examined in detail and at scale. They also represent a particular strength of the Cambridge English approach to testing, reflecting its long-term commitment to corpus-based methodologies and, more specifically, its extensive investment in developing and maintaining innovative collections of learner language; most notably,

1. Throughout, references to the A2 Key and the B1 Preliminary lists should be understood as encompassing the content for both the main exams and their variants for schools, reflecting the fact the lists for the two A2 Key exams are combined within a single document, while those for the B1 Preliminary exams are combined within another.

the *Cambridge Learner Corpus* and the *Cambridge English Profile Corpus* (Barker, Salamoura and Saville 2015, Harrison and Barker (Eds) 2015). Critically, however, and as has been argued elsewhere, such evidence does not negate the need for expert judgement (McCarthy and Carter 2003). Rather, it provides an empirical foundation on which assessment specialists can base their judgements, underpinning the quality of Cambridge English exams through drawing on multiple, complementary sources of evidence.

Updating the lists

Although both lists are generally subject to ongoing review, the immediate prompt for the present update was the specifications for the revised Key and Preliminary exams, due for first administration in 2020.² Accordingly, the purpose of this update was twofold. Firstly, to expand the breadth of lexis that would be available across the range of functions, genres and topics mandated by the revised specifications. Secondly, to help ensure that Key and Preliminary tasks continue to take account of contemporary language use; in particular, by adding words for new technology (e.g. 'app') and removing words for obsolete technology (e.g. 'floppy disk').

As detailed in the following section, a three-stage revision procedure was devised to meet these goals, in line with the longstanding Cambridge English approach of combining empirical data and expert judgement (Ball 2002, Street and Ingham 2007). This procedure was applied in two rounds, reflecting the wider demands of the overall revision process. The first rounds occurred in 2015 for the Key list and 2017 for the Preliminary list, so that item writers could begin producing the new tasks mandated by the revised specifications. The second round occurred in 2018, so that item writers and other assessment specialists could suggest new vocabulary that they felt would aid the item-writing process after extensive practical experience of producing the revised tasks.

The revision procedure

Stage 1: Identifying prospective vocabulary

For the initial stage, a list of prospective vocabulary for each exam was assembled from two sources: expert knowledge and the wordlists for two of the Young Learners exams.

Regarding the first source, item writers, item writer chairs, and other assessment specialists with particular expertise in A2 and B1 language suggested vocabulary items which they felt would expand the breadth of available lexis across the Key and Preliminary tasks that candidates would encounter when sitting the revised specifications. Secondly, the existing lists for two exams from Young Learners were reviewed: A1 Movers and A2 Flyers. Here, the principle was both to identify vocabulary that would expand the breadth of available lexis, and to establish a greater degree of conformity across the lists. Accordingly, these two Young Learners lists were selected on the basis of their targeting CEFR levels adjacent to those of Key and Preliminary, and hence constituting vocabulary that should be within the range of Key and Preliminary candidates.³ Note, however, that no systematic reconciliation exercise was conducted; that is, not all items in the Young Learners lists were selected as prospective items for the Key and Preliminary lists. This primarily reflects the differing candidature of the Young Learners exams, with the Key and Preliminary exams assessing vocabulary and contexts more suitable for older candidates (Papp and Rixon 2018). It also reflects the fact that each Cambridge English Qualification aims to assess learner development relative to that exam's targeted CEFR level. Accordingly, candidates for the higher-level Preliminary exam are expected to have already mastered the lower-level Key vocabulary and to access the Key list should they wish to review any vocabulary not also present on the Preliminary list.

2. Detailed information regarding the fully revised specifications can be found at: keyandpreliminary.cambridgeenglish.org

3. Since no general English vocabulary lists are provided for Cambridge English Qualifications above the B1 CEFR level, no B2 targeted list was available as a source of vocabulary for this level.

Stage 2: Collating the evidence

Once identified, the list of prospective words was sent for analysis by corpus specialists in the Research and Thought Leadership Group at Cambridge English. The aim of this stage was to provide empirical evidence regarding the vocabulary that Key and Preliminary candidates could reasonably be expected to know in contemporary English language contexts. With this in mind, it was decided to draw on three sources of evidence: two providing evidence of language use by English learners (the English Vocabulary Profile and the Cambridge Learner Corpus), and one providing evidence of language use in L1-speaking contexts (the Spoken British National Corpus 2014). These sources were analysed in the following sequence:

- a. Check whether each prospective vocabulary item already appears on the existing lists.
- b. Determine the CEFR level for each item according to the English Vocabulary Profile.
- c. Calculate the frequencies for each item as found in the Cambridge Learner Corpus.
- d. Calculate the frequencies for each item as found in the Spoken British National Corpus 2014.

Existing vocabulary lists

The initial step involved checking each prospective word against the existing Key and Preliminary lists.

First, a redundancy check was performed to see whether each prospective Key item was already present on the existing Key list, and each Preliminary item present on the existing Preliminary list. As with all steps, this procedure was performed using the specific sense and part-of-speech of the prospective vocabulary item. In the case of the Preliminary list, for example, the prospective verb 'doubt' was retained for further analysis since it was already present only as a noun, whilst the phrasal verb (to) 'deal with' in the sense of 'to deal with somebody' was also retained for further analysis since it was already present only in the sense of 'to deal with something'.

Second, each prospective item was reviewed against the vocabulary list for which they were not a candidate. The primary rationale here was to record the degree of potential overlap between lists that might result from adding any prospective words. As noted above, although some degree of overlap is appropriate, the Preliminary list is not intended to function as a superset of the Key list, since Preliminary candidates are expected to have already mastered the vocabulary appropriate to Key.

English Vocabulary Profile CEFR level

The next step involved looking up the CEFR level of each prospective item within the English Vocabulary Profile (EVP). Developed through a rigorous programme of lexicographic research, the EVP is a publicly available resource that sets out the typical words and phrases that English learners can be expected to know at CEFR Levels A1–C2 (Capel 2015).⁴ It thereby offered an invaluable resource for gauging whether prospective vocabulary items were appropriately pitched for the exam at hand.

In determining the appropriateness of prospective vocabulary, it was important to note that, whilst all Cambridge English Qualifications target a particular CEFR level, each exam is also designed to reliably assess candidates at adjacent levels using the Cambridge English Scale. Specifically, according to the strength of their performance, the Key exam allows for candidates to be assessed at A1, A2, or B1 level, whilst the Preliminary exam allows for candidates be assessed at A2, B1, or B2 level. This capacity for assessment at adjacent levels is reflected in the vocabulary which item writers have licence to draw on for each exam. Thus, for example, in the Key Listening paper, whilst the majority of the available grammar and lexis is specified so as to be characteristic of A2 language, the rubric in the more demanding Part 4 of the paper allows for up to 5% of the grammar and lexis drawn on to be characteristic of B1 language. The same is true for Preliminary, where the rubrics of Parts 2, 3 and 4 of the Reading paper, for example, are specified so as to include up to 5% B2 grammar and lexis.

Accordingly, this capacity to reliably assess Key and Preliminary candidates at adjacent levels was incorporated into the EVP review. Specifically, a prospective Key word was recommended for inclusion where identified by the EVP as characteristic of A1, A2, or B1 learner language. Conversely, a prospective Preliminary word was recommended for inclusion where identified as characteristic of A2, B1, or B2.

Frequencies in the Cambridge Learner Corpus

The third step involved determining the frequency of each prospective word at pre-determined CEFR levels using the Cambridge Learner Corpus (CLC). Developed over a period of two decades in conjunction with Cambridge University Press, the CLC comprises a systematic, regularly updated collection of written scripts produced by candidates of Cambridge English Qualifications from all over the world (Boyle and Booth 2000, Barker et al 2015).⁵ Currently standing at over 50 million words of learner writing, it constitutes a crucial source of evidence, not just of the characteristic language produced by English language learners at various CEFR levels, but the language specifically produced in response to Cambridge English Qualifications such as Key and Preliminary.

To enable direct comparison at the relevant CEFR levels, the relative frequencies of each prospective vocabulary item were calculated for those levels. Relative frequency is defined as the number of times an item occurs in a given subset of a corpus, expressed as a proportion of the total number of instances per million. For example, the relative frequency of 'password', a prospective word for the Preliminary list, has 0.75 instances per million words in the Key subset of the CLC but 0.94 instances per million in the Preliminary subset. Relative frequencies were used instead of raw frequencies in order to avoid distorting effects due to the differing sizes of each corpus subset, as is standard corpus practice (e.g. Leech, Rayson and Wilson 2001).

Regarding the specific CEFR levels checked, and in line with the principle of assessing Key and Preliminary candidates at adjacent levels, the relative frequency for each prospective word was calculated across three exams. These were: the exam for which it was a prospective word (i.e. Key or Preliminary), the exam targeting one CEFR level below, and the exam targeting one CEFR level above. For Key, this meant also calculating the relative frequencies for both Skills for Life Entry 1 (A1) and Preliminary (B1). For Preliminary, this meant also calculating the relative frequencies for Key (A2) and B2 First (B2).

Once their relative frequencies had been calculated, prospective vocabulary was recommended for inclusion as follows. For the Key list, vocabulary was recommended for inclusion where its relative frequencies were (a) low in Skills for Life Entry 1 but higher in Key and Preliminary, or (b) low in both Skills for Life Entry 1 and Key but had markedly increased so as to be more characteristic of Preliminary responses. For the Preliminary list, vocabulary was recommended for inclusion if their relative frequencies were (a) low in Key but higher in Preliminary and B2 First, or (b) low in both Key and Preliminary but had markedly increased so as to be more characteristic of B2 First responses.

Frequencies in the Spoken British National Corpus 2014

The final source of evidence was the Spoken British National Corpus 2014 (Spoken BNC2014). This is an 11.5-million-word corpus of spontaneous conversations between L1 speakers of British English, collected between 2012 and 2016 by Lancaster University and Cambridge University Press (Love, Dembry, Hardie, Brezina and McEnery 2017).⁶ Constituting the most up-to-date source of information on contemporary L1 English, it offers strong evidence regarding the vocabulary that L2 learners can currently expect to encounter in L1-speaking contexts.

5. The CLC is available for research purposes, and researchers interested in accessing the CLC are invited to do so via the following link: languageresearch.cambridge.org/academic-research-request-form

6. As with the EVP, the Spoken BNC2014 is publicly available, and can be accessed free of charge at: corpora.lancs.ac.uk/bnc2014

Using the Spoken BNC2014, prospective vocabulary items were evaluated as follows. Firstly, items representing technological phenomena were recommended for consideration where present with a corpus frequency equal to or more than one instance per million words. Secondly, it was recommended that items not representing technological phenomena should be excluded from the final vocabulary list where present with a frequency of less than one instance per million words. In both cases, the principle was to avoid burdening candidates with vocabulary that they are unlikely to require, whilst highlighting vocabulary of contemporary salience that might not yet have translated through to learner responses; at least as evidenced in the CLC at the point of revision.

Stage 3: Reviewing the evidence

Once collated, the resulting evidence base was passed to a team of assessment specialists at Cambridge English, who met to evaluate the evidence and make any final decisions. These specialists comprised Assessment Managers currently working on the various Key and Preliminary products, as well as specialists with wider expertise in learner language at A2 or B1 levels.

Throughout this stage, the default decision-making principle was to accept all prospective items where the empirical analysis had recommended them for inclusion on the basis of (a) their CEFR level in the EVP or (b) their CLC frequencies. Thus, of the prospective Key vocabulary, the adjective 'delicious' was accepted, for example, since it is identified by the EVP as characteristic of B1 language, and its CLC frequencies evidence a substantive increase in Key and Preliminary writing. However, the adjective 'mean' was rejected, since the EVP characterises it as B2 language, and its CLC frequencies evidence no clear increase in Key or Preliminary writing. Conversely, of the prospective Preliminary vocabulary, the adjective 'enjoyable' was accepted, since it is identified by the EVP as characteristic of B1 language, and its CLC frequencies evidence a marked spike in the writing of B2 First candidates. However, the adjective 'suited' was rejected, since the EVP identifies it as characteristic of C1 language, and the CLC frequencies indicate that it is effectively absent from Preliminary or B2 First writing.

Nevertheless, reflecting the wider principle of combining empirical data and expert judgement, prospective vocabulary was also accepted where a clear justification could be made for its inclusion. For example, the noun 'superhero' would otherwise have been rejected as a prospective word for the Preliminary list, due to its very low CLC frequencies and its complete absence from the EVP. However, it was decided to include this word as it was deemed to be an internationally recognised word with a high contemporary salience that would support item writers in creating engaging tasks for the Preliminary candidature.

As a final step, the resulting vocabulary lists were also reviewed to eliminate vocabulary that was obsolete from the specific perspective of Key and Preliminary candidates, such as 'floppy disk'. Again, this step underlines the value of combining empirical analysis and expert judgement, since corpus-based frequencies may not easily distinguish generally obsolete words from those that may still be useful to Key and Preliminary candidates. For example, both 'floppy disk' and 'CD player' have relatively low frequencies according to the Spoken BNC2014 (respectively, 0.44 per million and 0.96 per million). Nevertheless, there was a clear argument for retaining 'CD player' in the Key vocabulary list, since CD players are still widely used in various educational contexts, and indeed Key Listening tests are still also made available on CD. Hence, this is a word that Key candidates may well need to know.

The revised 2020 vocabulary lists

The fully revised vocabulary lists were published in 2018 ahead of the revised 2020 Key and Preliminary specifications, and remain freely available on the Cambridge English website.^{7,8} In terms of overall vocabulary size, the 2020 Key Vocabulary List now contains 1,599 words and phrases, representing an increase of 76 items (or 4.99%). Conversely, the 2020 Preliminary Vocabulary List now contains 3,046 words and phrases, representing an increase of 70 items (or 2.35%). These are relatively small increases, representing a balance between (a) providing item writers with enough lexical variety to design a range of appropriate Key and Preliminary tasks, whilst (b) providing Key and Preliminary candidates with an expanded set of vocabulary that they can target without being unduly burdened.

As is standard practice, the revised Key and Preliminary vocabulary lists will be subject to ongoing review and updates by assessment specialists at Cambridge English, taking advantage of contemporary developments in language testing and corpus-based research. Indeed, a further review is currently planned for 2021, in order to address any improvements that are identified following the release of the revised specifications. A likely focus of updates in the longer term will be the salience of formulaic language such as collocations and lexical bundles, reflecting the increasing awareness and understanding of specific word combinations as a key dimension of L2 proficiency (Gyllstad 2013, Henriksen 2013, Siyanova-Chanturia and Pellicer-Sánchez (Eds) 2019). Whatever the specific focus and outcomes, however, the core purpose of future updates will remain the same: to ensure the ongoing currency of the Key and Preliminary exams so as to maximally support the evolving learning and assessment needs of A2 and B1 learners.

References

- Ball, F (2002) Developing wordlists for BEC, *Research Notes* 8, 10–13.
- Barker F, Salamoura, A and Saville, N (2015) Learner corpora and language testing, in Granger, S, Gilquin, G and Meunier, F (Eds) *The Cambridge Handbook of Learner Corpus Research*, Cambridge: Cambridge University Press, 511–533.
- Boyle, A and Booth, D (2000) The UCLES/CUP Learner Corpus, *Research Notes* 1, 10.
- Capel, A (2015) The English Vocabulary Profile, in Harrison, J and Barker, F (Eds) (2015) *English Profile in Practice*, English Profile Studies volume 5, Cambridge: UCLES/Cambridge University Press, 9–27.
- Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Cambridge: Cambridge University Press.
- Council of Europe (2018) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors*, Strasbourg: Council of Europe.
- Docherty, C (2015) Revising the Use of English component in FCE and CAE, *Research Notes* 62, 15–20.
- Gyllstad, H (2013) Looking at L2 vocabulary knowledge dimensions from an assessment perspective: Challenges and potential solutions, in Bardel, C, Lindqvist, C and Laufer, B (Eds) *L2 Vocabulary Acquisition, Knowledge and Use: New Perspectives on Assessment and Corpus Analysis*, Eurosla Monograph Series volume 2, 11–28.
- Harrison, J and Barker, F (Eds) (2015) *English Profile in Practice*, English Profile Studies volume 5, Cambridge: UCLES/Cambridge University Press.
- Henriksen, B (2013) Research on L2 learners' collocational competence and development: A progress report, in Bardel, C, Lindqvist, C and Laufer, B (Eds) *L2 Vocabulary Acquisition, Knowledge and Use: New Perspectives on Assessment and Corpus Analysis*, Eurosla Monograph Series volume 2, 29–56.
- Leech, G, Rayson, P and Wilson, A (2001) *Word Frequencies in Written and Spoken English: Based on the British National Corpus*, London: Longman.

7. www.cambridgeenglish.org/images/506886-a2-Key-2020-vocabulary-list.pdf

8. www.cambridgeenglish.org/images/506887-b1-Preliminary-2020-vocabulary-list.pdf

- Love, R, Dembry, C, Hardie, A, Brezina, V and McEnery, T (2017) The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations, *International Journal of Corpus Linguistics* 22(3), 319–344.
- McCarthy, M and Carter, R (2003) What constitutes a basic spoken vocabulary? *Research Notes* 13, 5–7.
- Papp, S and Rixon, S (2018) *Examining Young Learners: Research and practice in assessing the English of school-age learners*, Studies in Language Testing volume 47, Cambridge: UCLES/Cambridge University Press.
- Siyanova-Chanturia, A and Pellicer-Sánchez, A (Eds) (2019) *Understanding Formulaic Language: A Second Language Acquisition Perspective*, Abingdon: Routledge.
- Street, J and Ingham, K (2007) Publishing vocabulary lists for BEC Preliminary, PET and KET exams, *Research Notes* 27, 4–7.
- Van Ek, J A and Trim, J L M (1991a) *Threshold 1990*, Strasbourg: Council of Europe.
- Van Ek, J A and Trim, J L M (1991b) *Waystage 1990*, Strasbourg: Council of Europe.

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

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.

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).

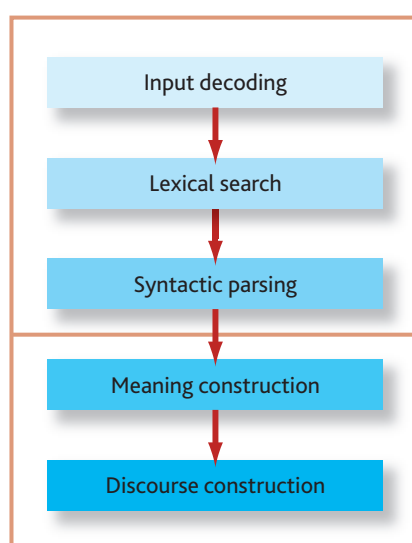


Figure 1: Cognitive processing model adapted from Geranpayeh and Taylor (Eds) 2013

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.

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.

Table 1: Revised Key Listening test format

<i>Current: Key Listening (until end 2019)</i>		<i>Trialing version 1: Key Listening</i>		<i>Trialing version 2: Key Listening Revised test format (from January 2020)</i>	
<i>Timing</i>	<i>Content</i>	<i>Timing</i>	<i>Content</i>	<i>Timing</i>	<i>Content</i>
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.

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%

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

<i>Language</i>	<i>Number</i>	<i>%</i>
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).

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.

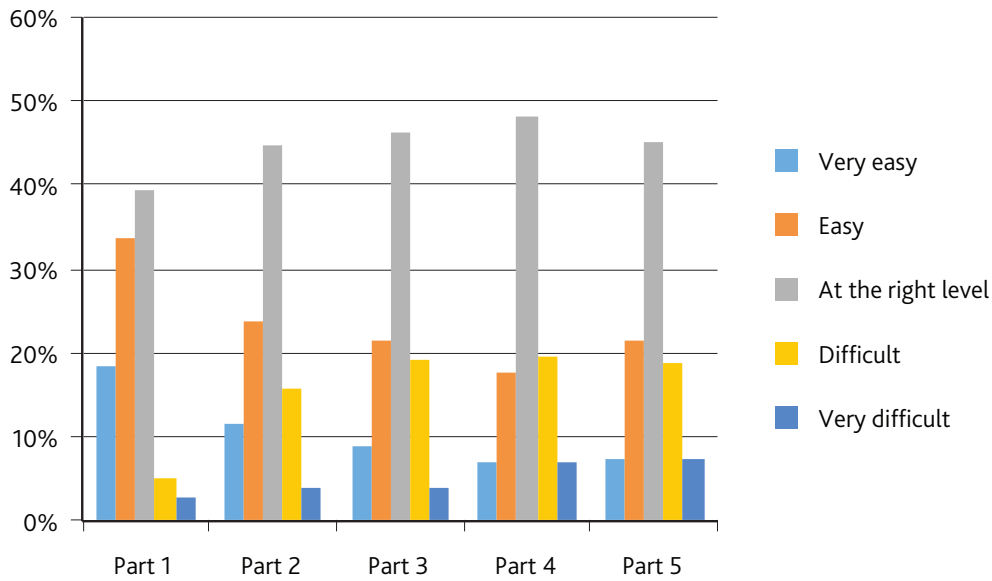


Figure 2: Candidates' perceptions of the difficulty of the test by part

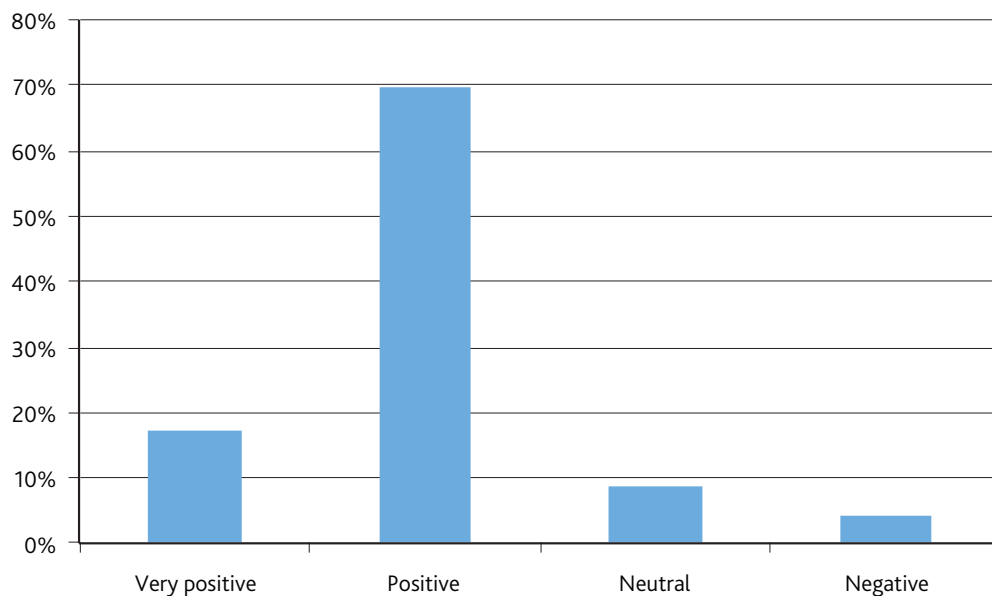


Figure 3: Teachers' perceptions of the revised test

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, automatising 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.

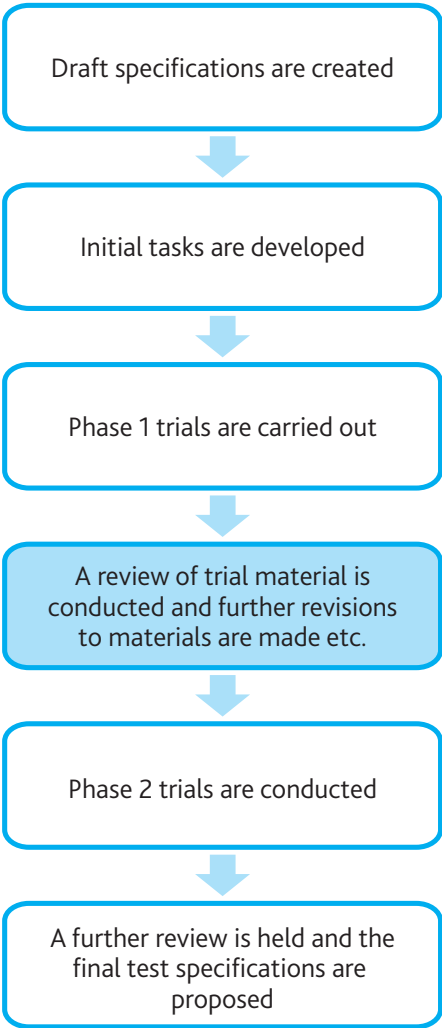


Figure 4: Revised test trialling process

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) (2013) *Examining Listening: Research and Practice in Assessing Second Language Listening*, Studies in Language Testing volume 35, Cambridge: UCLES/Cambridge University Press.

Revising the A2 Key and B1 Preliminary Reading exam

Marianne Pickles Assessment, Cambridge Assessment English

Tony Clark Research and Thought Leadership, Cambridge Assessment English

Mark Elliott Validation and Data Services, Cambridge Assessment English

Introduction

At Cambridge Assessment English, we regularly review the performance of our exams to ensure that they continue to be fit for purpose. In the years since the last revision of A2 Key and B1 Preliminary in 2004, there have been changes to our understanding of and beliefs about the General English proficiency construct at the A2 and B1 levels which we wished to ensure were reflected in the new specifications of the exams during this revision process. The introduction of upwards certification, whereby very high-performing candidates can be granted a certificate at one Common European Framework of Reference for Languages (CEFR, Council of Europe 2001) level above that of the target level of the exam they are taking, means it is necessary for each exam to feature elements of the descriptors for the level above. There has also been a move towards greater standardisation, where appropriate, from one Cambridge English Qualification to the next in order to provide a greater sense of continuity between these exams for the candidates and their teachers. This article will discuss the changes made to the Reading components of the A2 Key and B1 Preliminary exams and the rationale behind these decisions in light of these considerations.¹

It should be noted that while the Reading and Writing components of Preliminary have now been split into two separate papers, there continues to be a combined Reading and Writing paper for Key. For discussion of the writing tasks for Key and Preliminary, see Panagiotopoulou, Lambie and Cheung, this issue.

Refining Reading test constructs

The first language (L1) reading construct, shown in Figure 1, details the types of reading that may be used by L1 readers in the left-hand column. Careful reading involves extracting the whole meaning, while expeditious reading involves reading quickly and selectively (Khalifa and Weir 2009, Urquhart and Weir 1998). Local reading refers to understanding individual sentences, whereas global reading refers to comprehension across sentences or of the overall text. Whatever combination of careful or expeditious and local or global is chosen will depend on the purpose for reading. For example, searching for a particular word in the dictionary will involve scanning, a type of expeditious local reading, while reading the definition will involve careful local reading. Even at A2 level, students have been found to be able to make use of both the expeditious and careful reading types (Pickles 2018). For the revised A2 Key and B1 Preliminary exams, we therefore considered it important to include at least one task that could encourage expeditious reading. The 2004 test specifications for A2 Key did not explicitly feature such a task, but by contrast, the B1 Preliminary test did – the true–false task in Part 3.

1. Where reference is made to 'A2 Key', this should be read as inclusive of A2 Key for Schools, and likewise the term 'B1 Preliminary' within this article encompasses the standard and the variants for schools. These exams were previously known as *Cambridge English: Key* and *Cambridge English: Preliminary*.

However, several issues with the task were identified:

- the response format is naturally amenable to guessing since there is a 50% chance of answering correctly by responding at random; this was reflected in the performance of the items, which typically exhibited poorer measurement characteristics than those of other tasks
- perhaps also as a result of the binary response format, the task itself was rather straightforward, so generating items of the required level of difficulty often resulted in the use of texts which were effectively above the level of the exam, resulting in an imbalanced task.

For these reasons it was considered necessary to revisit the expeditious reading coverage for both sets of revised specifications.

The central column entries in Figure 1 show the cognitive processes involved in L1 reading, which increase in complexity as the list ascends. 'Inferencing', the first of the higher-level processes, is generally considered to come into play at Level B2 and above (Khalifa and Weir 2009). It is the lower-level processes that are relevant for Levels A2 and B1. In the 2004 specifications for A2 Key, 'lexical access' and 'syntactic parsing' are tested through multiple-choice cloze tasks, while 'establishing propositional meaning' is tested through tasks which test the comprehension of both shorter and longer texts. A2 Key also tested syntactic parsing through an open cloze task, but this task type was absent from B1 Preliminary. It was important to address this by adding such a task to the revised B1 Preliminary exam while retaining the existing testing focuses.

Reviewing and revising the tasks

Bearing in mind the priorities described in the previous section, the 2004 specification task types for both A2 Key and B1 Preliminary were reviewed in order to judge to what extent they contributed towards the exams being fit for purpose. Subject specialists with many years of expertise in writing for and reviewing these exams were commissioned to write reports on this, and our internal research department also ran analyses to evaluate the tests' construct coverage and measurement characteristics.

In particular, Structural Equation Modelling (SEM) was conducted to evaluate the consistency of the tasks within and across components of B1 Preliminary (Elliott, Benjamin and Docherty 2014b), which revealed that Writing Part 1 consistently loaded on Reading rather than Writing. This finding is not entirely surprising since it is a sentence transformation task, a Use of English task type resembling those in B2 First, C1 Advanced and C2 Proficiency, although the fact that they typically involve a single word makes them closer to open cloze items. The Use of English construct overlaps with Reading in that it relies on lower-level reading cognitive processes, and at B1 level, where there are few higher-level cognitive processes involved, the two constructs are close enough to be considered unidimensional. The decision was made to move the task into Reading but to convert it into a standard open cloze task with a single extended text rather than discrete sentences, which brings the added advantage of consistency across all five levels (A2 Key, B2 First, C1 Advanced and C2 Proficiency all feature such a task).

Differential Item Functioning (DIF) analysis, which is a statistical technique used to identify items which behave differently (i.e. have a different level of difficulty) for different defined groups (e.g. male and female) when ability is accounted for, was conducted to investigate potential bias in particular tasks in A2 Key (Elliott, Benjamin and Docherty 2014a). A2 Key Reading and Writing Part 6 – a vocabulary/spelling task – was found to exhibit DIF across different first languages in some cases. Since there were also construct-related issues with the task (discussed in the next section), the decision was made to remove it from the test.

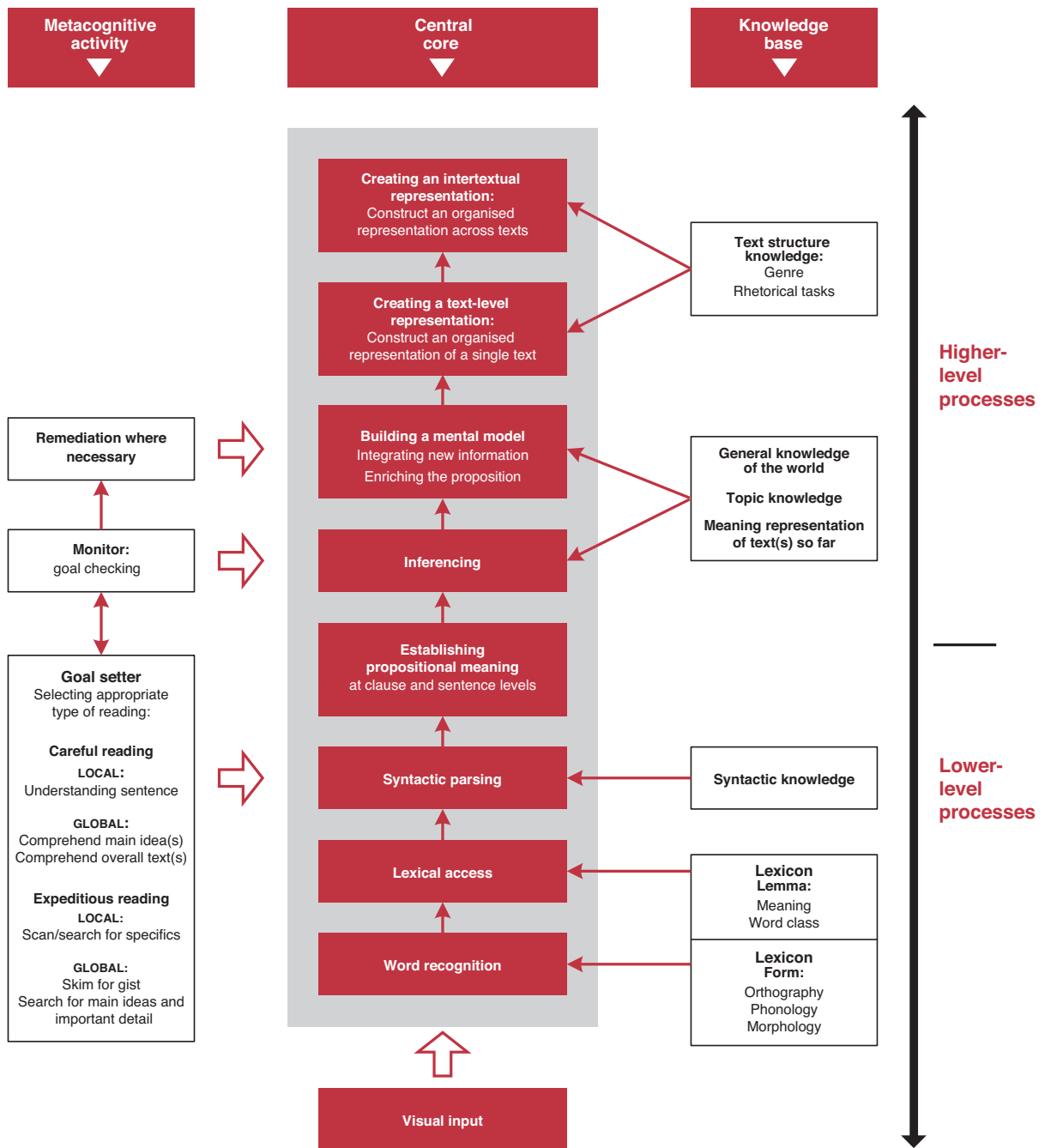


Figure 1: Khalifa and Weir's (2009) model of reading
 Source: Adapted from Khalifa and Weir (2009:43) (Brunfaut and McCray 2015:7)

An overview of the content of the revised A2 Key and B1 Preliminary exams is shown in Tables 1 and 2 respectively. Table 3 lists the tasks from the 2004 specifications which were retired and outlines the reasons for this. A more detailed account follows in the next section.

Table 1: The revised A2 Key Reading paper

<i>Exam part (items; marks)</i>	<i>Task type</i>	<i>Task focus</i>	<i>Notes</i>
Part 1 (6 items; 6 marks)	3-option multiple choice (discrete)	Reading comprehension for detail and main ideas at word, phrase and sentence level.	New. Same task type as B1 Preliminary Part 1.
Part 2 (7 items; 7 marks)	3-option multiple matching	Reading comprehension for specific information at sentence and text level; expeditious reading and careful reading.	New. Adapted from a task type option for A2 Key 2004 specifications Part 4.
Part 3 (5 items; 5 marks)	3-option multiple choice	Reading comprehension for detail and main ideas at sentence and text level; careful reading.	New. Adapted from a task type option for A2 Key 2004 specifications Part 4.
Part 4 (6 items; 6 marks)	3-option multiple-choice cloze	Lexical and lexico-grammatical words in context.	Modified. Given a more lexical focus.
Part 5 (6 items; 6 marks)	Open cloze	Grammatical words in context.	Modified. Number of items reduced.

Note: Writing is part of the same exam for A2 Key and Parts 6 and 7 are writing tasks (see Panagiotopoulou, Lambie and Cheung, this issue).

Table 2: The revised B1 Preliminary Reading paper

<i>Exam part (items; marks)</i>	<i>Task type</i>	<i>Task focus</i>	<i>Notes</i>
Part 1 (5 items; 5 marks)	3-option multiple choice (discrete)	Reading comprehension for detail and main ideas at word, phrase and sentence level.	As B1 Preliminary 2004 specifications.
Part 2 (5 items; 5 marks)	8-option multiple matching	Reading comprehension for specific information at sentence and text level; expeditious reading and careful reading.	As B1 Preliminary 2004 specifications.
Part 3 (5 items; 5 marks)	4-option multiple-choice reading	Reading comprehension for detail and main ideas at sentence and text level; careful.	As Part 4 in B1 Preliminary 2004 specifications.
Part 4 (5 items; 5 marks)	Gapped text	Reading comprehension for main ideas and understanding coherence and cohesion at text level.	New. Added to cover 'building a mental model'.
Part 5 (6 items; 6 marks)	4-option multiple-choice cloze	Lexical and lexico-grammatical words in context.	Modified. Given a more lexical focus.
Part 6 (6 items; 6 marks)	Open cloze	Grammatical words in context.	New. Added for construct coverage.

Table 3: Retired tasks from the A2 Key and B1 Preliminary 2004 specifications

<i>Exam and part</i>	<i>Task type</i>	<i>Task focus</i>	<i>Notes</i>
A2 Key Part 1	Matching	Reading comprehension for detail at sentence level.	Some items were interdependent.
A2 Key Part 3A	3-option multiple choice (discrete)	Reading comprehension and pragmatic competence.	Now tested more authentically in the Speaking paper.
A2 Key Part 3B	Gapped text	Reading comprehension and coherence and cohesion in a dialogue.	Now considered inauthentic for the Reading paper.
A2 Key Part 4	Right–Wrong–Doesn't Say (R–W–DS)	Reading comprehension for detailed understanding and main ideas at text level.	R–W–DS no longer considered cognitively appropriate for A2 readers.
A2 Key Part 6	Spelling and definitions	Vocabulary knowledge and correct spelling.	Now tested more authentically in the Writing paper.
A2 Key Part 8	Information transfer	Identifying specific information from a pair of short texts.	Some items were interdependent.
B1 Preliminary Part 3	True–False	Reading comprehension for specific information; expeditious reading.	Binary tasks no longer desired.

Changes to A2 Key Reading – more reading comprehension

As explained in the previous section, it was considered important for the A2 Key exam to feature a task which aimed to elicit expeditious reading of a longer text as well as one that was more focused on careful reading. The multiple-matching and multiple-choice variants of the A2 Key 2004 specifications Part 4 tasks were considered suitable task types for these respective purposes and the first trial tests featured versions of each of these. By contrast, the Right–Wrong–Doesn't Say (R–W–DS) task type was retired on the basis that there was some debate around whether it is appropriate, from a cognitive perspective, to ask A2 candidates to conceptualise the idea of a text 'not saying' something, as opposed to saying that something is false. Based on the statistics for this task type, the candidates did not appear to have difficulty in completing it, but the question over its cognitive validity, as well as the lack of any comparable task among the rest of the Cambridge English Qualifications, led to its being retired.

The same desire to include more reading comprehension tasks led to the 2004 specifications Part 1, a matching task, being replaced with a lower-level equivalent of the B1 Preliminary 2004 specifications Part 1, a series of discrete 3-option multiple-choice tasks.

After these changes were made, the revised A2 Key Reading test contained 18 reading comprehension items as opposed to eight. However, the inclusion of these denser reading tasks also meant an increase in reading load. For the benefit of the candidates, their exam day experience, and the avoidance of fatigue, it was considered important that the revised exam take a similar amount of time to complete as its predecessor. It was therefore necessary to consider whether any of the remaining 2004 specifications tasks were surplus to requirements in terms of representing the target construct.

Changes to A2 Key Reading – avoiding duplication

It was noted that the 2004 specifications of A2 Key featured two multiple-choice cloze tasks – the 5-item Part 2 and the 10-item Part 5. The former had more of a focus on lexis while the latter had a more grammatical focus. In order to avoid the inclusion of extraneous items where an aspect of the construct was being covered thoroughly

in multiple places, it was decided that these tasks would be combined. The revised task became the new Part 4, which is a 6-item multiple-choice cloze task in a text format. Four or five of the items in each task are to focus primarily on lexis and one or two items may have a more grammatical focus. The open cloze task was retained with fewer items focusing primarily on grammar. The revisions to these tasks ensured that construct coverage was maintained with less redundancy.

Changes to A2 Key Reading – construct relevance

There were further tasks within the A2 Key 2004 Reading specifications which added to the length of the exam, but which seemed to target a slightly broader definition of the reading construct than was now conceptualised. Part 3A was a series of five multiple-choice discrete items where students were required to choose the pragmatically appropriate response. Similarly, Part 3B featured a written conversation between two people from which one interlocutor's turns had been removed and the candidates had to put them back into the text in the right order. Something these tasks had in common was that they dealt with issues relating to interaction, but in a purely receptive context. The aspects of language that these tasks were testing were felt to sit more logically within the Speaking component of the exam and in the case of Part 3B, the notion of presenting students with the script of a spoken conversation now seemed somewhat less authentic both interactionally and situationally.

Similarly, the testing of vocabulary definitions and correct spelling, which occurred in Part 6, seemed better suited to the authentic context of writing a text than isolated as a separate task. Furthermore, this particular task format was found to be prone to a degree of DIF, which could be a source of bias, and the dual focus on vocabulary knowledge and spelling (particularly precise spelling, which is not entirely consistent with the performance descriptors for A2) introduced a degree of construct ambiguity. Part 8, which tested information transfer, was also identified for removal from the specifications. It could suffer from interdependency and, as such, the tasks were very hard to write to the extent that it would have bordered on impractical to keep it. Additionally, the candidates found the tasks so easy that Part 8 frequently failed to meet the statistical requirements after pretesting, resulting in a great deal of rewriting, retesting and wastage. On the basis of these considerations, the task was retired.

The changes described above meant that we were able to ensure that each item in the test was more focused on truly testing reading.

Changes to B1 Preliminary Reading

Just as for A2 Key, experienced subject specialists were commissioned to write reports about the B1 Preliminary exam and provide any recommendations for changes alongside our own research department conducting analyses into task performance. The findings revealed that a high proportion of the existing task types within B1 Preliminary were performing as desired. Two changes were considered necessary. The first of these was to remove the true–false task in Part 3. Its binary nature meant it was too amenable to guessing and there was often an imbalance between the difficulty of the task and the complexity of the text.

The other change made was to incorporate 'building a mental model' into the reading construct covered within the test. Building a mental model, even at paragraph level as this task requires, is arguably an upper B1, even a B2 reading ability, and it was felt that the inclusion of this task, if trialling proved successful, may help in discriminating between candidates for the purposes of upwards certification. The task trialled well and it was confirmed that the task was suitable for use at B1 level, provided that the connections within and across sentences were clear and the text followed a simple temporal sequence. A further benefit of the task type was that it tested understanding at a global level, across sentences rather than simply within them. Our subject specialists additionally noted that the inclusion of this task would create greater continuity between B1 Preliminary and B2 First. These two areas constitute the extent of the changes to B1 Preliminary Reading.

Creating continuity between the exams

In addition to the considerations detailed above with regard to the test construct, the revision of A2 Key and B1 Preliminary was viewed as an ideal opportunity to revisit the structure of the exams and enhance a sense of continuity between the levels. In terms of the 2004 specifications, both A2 Key and B1 Preliminary featured task types that were unique to each exam. For instance, A2 Key Reading featured a task about definitions and spelling (Part 6) and B1 Preliminary Reading contained a true–false task (Part 3). In order to help learners and teachers by reducing the variety of task types where possible and aligning the structure of the exams, both of these task types were removed and a number of additional changes were made. It has been noted that although students' preferences for task types are often diverse, certain formats are believed to cause less anxiety than others, which are usually reflective of their preferred learning style (Birenbaum 2007). However, the effect of task *familiarity* on test takers was particularly relevant when revising these exams. Candidate performance may be improved if familiar task types are encountered (Révész and ZhaoHong 2006), which therefore had to be considered during the exam revision process.

The order of the revised A2 Key exam was modelled after that of B1 Preliminary, with Parts 1 and 2 mirroring each other. This resulted in both exams beginning with multiple-choice discrete tasks followed by a multiple-matching task geared towards eliciting expeditious reading. In both exams, Part 3 is a multiple-choice reading comprehension task with a longer text. B1 Preliminary then features a gapped text task for which there is no A2 equivalent. Both Reading exams are concluded with a multiple-choice cloze followed by an open cloze. The structures of the exams are therefore now aligned. At one stage, consideration was given to attempting to replicate the order of tasks in B2 First, C1 Advanced and C2 Proficiency. This would have meant beginning the exam with the multiple-choice cloze task. However, it was felt that the discrete reading comprehension tasks would make for a more appropriate introduction to the exam for these lower-level learners, and that it would be beneficial for those teachers and students already familiar with the structure of B1 Preliminary.

Trialling the new specifications

There were two trials for A2 Key Reading. Between Trials 1 and 2, changes were made to the format of the tasks, the rubrics for the tasks, and the number of items in each task.

Revised A2 Key trial outcomes

The second trial of the revised A2 Key Reading paper in August 2016 formed part of a report which explored the validity and reliability of the revised Listening, Reading and Writing components, in addition to the suitability of time provisions for candidates (Vidaković 2018a). Although Reading and Writing sections are combined in one paper in A2 Key (see Table 1), it was straightforward to separate findings related to reading constructs as required. The performances of a sample of candidates (330 from 12 test centres) were analysed, and surveys about their perspectives on the test conducted (294 responses). A small sample of teachers and invigilators (N=9) were also surveyed, to explore perspectives on the revised test. In terms of participant characteristics, most candidates were of the target age group of the live exam for schools (11–14 years old). Regarding CEFR level, most participants (63%) were of A2 (the target level) and 25% were one band below at A1, which accurately reflected the expected live candidature.

Results of statistical item analysis indicated that there were no significant issues with the difficulty of the revised Reading test. Supporting this data, the candidates' survey responses indicated that most (90%) agreed that the test allowed them to demonstrate their reading ability. The majority of participants (67%) perceived the difficulty of the paper as they would have expected it to be. The time taken by candidates was also compared to their reported CEFR level, indicating that timing was appropriately accounted for in the new test version. Results supported the finding that timing was not a considerable issue. In the survey, 81% of respondents felt that they had enough time to complete the test. Finally, the teacher and invigilator survey responses were also largely positive, further supporting the overall outcome of the trial.

Revised B1 Preliminary trial outcomes

The second round of trialling was also conducted for B1 Preliminary between August and September 2016, to further investigate the validity and reliability of the revised test, and the adequacy of time allowed (Vidaković 2018b). For the Reading part of the exam, the performances of 443 candidates were analysed. In terms of CEFR level, test takers were broadly similar to those expected to take part in the live candidature – although some were slightly weaker. 40% of participants were B1 (target level) and 40% at A2 (one level below the target band). 285 candidates also completed a survey about their experiences of the test, including 77% aged 15 years old or below, which reflects the younger target age group for B1 Preliminary, and compared favourably to the live candidature. Other characteristics of the trial participants were reflective of actual candidates; most were preparing to take the real test (74%), had not taken it before (88%) and intended to do so within the next 12 months (90%). In addition to the candidates, a group of 11 teachers and invigilators also completed a survey about their perceptions of the test.

The results of the statistical item analysis highlighted several potentially problematic items across the tasks. However, these results were found to be due to issues with individual items rather than the tasks themselves. Part 6 (the new open-cloze section) had a higher than acceptable range of item difficulties for B1 Preliminary. Nonetheless, erratic item difficulties such as these do not appear to suggest a systemic problem, and would be possible to address through familiarisation and training of item writers. However, indications were that – overall, at least – the timings were appropriate. Candidates taking more than 45 minutes to complete the test were 4.4% at B2, 14% at B1, 29.2% at A2 and 31.8% for below A2. As the expected proficiency level of most test takers was B1/A2, the timings were demonstrated to be appropriate.

Candidates' survey feedback about the test was largely positive, with 89% agreeing that it was a good means of demonstrating their reading ability, and 92% indicating that the instructions were clear. The perceived difficulty of the paper revealed mixed responses, but it was apparent that the two new tasks (Parts 4 and 6, the gapped text and open cloze, respectively) were viewed as the most challenging. Part 4 (which involved removed sentences) was believed to be particularly difficult, with 36% of candidates reporting it as difficult or very difficult. This finding supported the earlier item-level statistical analysis, which found Part 4 to be the most complex. This may have been as it was a new and unfamiliar task type at the level, as will be subsequently discussed. However, the overall difficulty of the revised test was as expected for the majority of test takers (67%). Regarding timing, most participants felt this was 'OK' (76%). Teachers and invigilators were also mostly enthusiastic about the revised Reading test, with eight (N=11) identifying as 'positive' or 'very positive', and two were neutral. All of these participants agreed that the tasks were important and worth students' time to prepare for; nine (N=11) believed that they were motivating and that the instructions were clear. With regard to timing, the teachers and invigilators mostly agreed with the candidates, as most (seven, N=11) reported the overall timing to be 'OK'. They also supported the findings from candidates' scores and survey responses reported above. This particularly focused on the new Part 4 as being too difficult (six, N=11, reported agreement of its appropriateness) for students based on its unfamiliarity, in addition to the new Part 6, albeit to a lesser extent (eight, N=11, agreed that it was at the

appropriate level). Finally, as was the case with the A2 Key revision trial, several teachers and invigilators also commented that rudimentary practical aspects of the test such as shading the boxes might have been challenging for some learners. They also recommended a wider range of topics available to candidates in order to increase fairness. Overall, however, their perspectives on the test were largely favourable.

Conclusion

The revisions of A2 Key and B1 Preliminary have improved each exam in several ways, each of which is evidently advantageous to the test taker. Rigorous trialling of these revisions has indicated that although making such changes can present certain unforeseen challenges at times, the overall outcome in each case was generally highly positive. Including new task types allowed both exams to better reflect contemporary understanding of the reading construct, for example. Adding a task (in the A2 Key exam) which encouraged expeditious reading had not been done previously, and removing the binary expeditious task in B1 Preliminary was equally important to maintain fitness for purpose in a contemporary assessment context. Changing the order of the parts (A2 Key Parts 1 and 2, for example) to standardise the structure of the two exams was another principal objective. However, in each case the difficulty of the items and the overall timing of the test had to be investigated, if the impact of the revisions was to be thoroughly understood. Furthermore, it had to be determined if this would affect stakeholder experiences; exploring candidate and teacher perspectives on these modifications was an essential part of understanding the value of the changes.

In summary, the latest trial results for both revised exams were largely positive. The timing of each test was found to be unproblematic, and candidates generally appeared to have had enough time to complete the exam. Respondents mostly felt that instructions were clear, the tests allowed candidates to demonstrate their reading ability, and that the overall difficulty was as expected in each case. In terms of specific tasks, difficulty was not a significant issue with A2 Key, and each of the revised items performed as expected. For B1 Preliminary, there were a few relatively minor issues highlighted by the later trial. In particular, the two new tasks (Parts 4 and 6) contained items that were found to be more difficult than others, as explained above. However, this did appear to be as a result of a lack of familiarity with the individual items in question, and it was concluded that item writer training could be used to overcome this. The latest revisions of both the A2 Key and B1 Preliminary exams – and the subsequent trials to determine their impact – have evidently produced largely favourable results, which indicate that the revisions to the Reading papers will be beneficial to candidates and teachers.

References

- Birenbaum, M (2007) Assessment and instruction preferences and their relationship with test anxiety and learning strategies, *Higher Education* 53 (6), 749–768.
- Brunfaut, T and McCray, G (2015) *Looking into test-takers' cognitive processes whilst completing reading tasks: a mixed-method eye-tracking and stimulated recall study*, ARAGs Research Reports Online, volume AR/2015/001, London: The British Council, available online: www.britishcouncil.org/sites/default/files/brunfaut_and_mccray_report_final.pdf
- Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Cambridge: Cambridge University Press.
- Elliott, M, Benjamin, T and Docherty, C (2014a) *Cambridge English: Key/Key for Schools DIF analysis*, Cambridge Assessment English Internal Validation Report 1570.
- Elliott, M, Benjamin, T and Docherty, C (2014b) *Cambridge English: Preliminary/Preliminary for Schools Construct Investigation (SEM)*, Cambridge Assessment English Internal Validation Report 1627.

- Khalifa, H and Weir, C J (2009) *Examining Reading: Research and Practice in Assessing Second Language Reading*, Studies in Language Testing volume 29, Cambridge: UCLES/Cambridge University Press.
- Pickles, M (2018) *Types of reading in an A2 level test of English as a second language: A mixed methods investigation into Spanish-L1 ESL learners' completion of multiple-matching reading tasks*, Master's dissertation.
- Révész, A and ZhaoHong, H (2006) Task content familiarity, task type and efficacy of recasts, *Language Awareness* 15 (3), 160–179.
- Urquhart, S and Weir, C (1998) *Reading in a Second Language: Process, Product and Practice*, Essex: Longman.
- Vidaković, I (2018a) *Cambridge English: Key (for Schools) Revision Across Skills. Sep 2016 Trial*, Cambridge Assessment English Internal Validation Report 1720.
- Vidaković, I (2018b) *Cambridge English: Preliminary (for Schools) Revision Across Skills. Sep 2016 Trial*, Cambridge Assessment English Internal Validation Report 1721.

Revising the A2 Key and B1 Preliminary Speaking exam

Kathryn Davies Assessment, Cambridge Assessment English

Nick Glasson Assessment, Cambridge Assessment English

Introduction

Cambridge Assessment English periodically reviews all of its assessments to guarantee fitness for purpose. The review of A2 Key and B1 Preliminary was carried out to ensure that these exams remain relevant to test users' evolving needs. Further aims were to facilitate progress up the 'proficiency ladder' through better alignment with exams at higher levels, improve alignment with the Common European Framework of Reference for Languages (CEFR, Council of Europe 2001), and have a positive impact on teaching and learning.¹

Cambridge English has an established process for exam revision, and for Speaking components it is summarised in Figure 1.

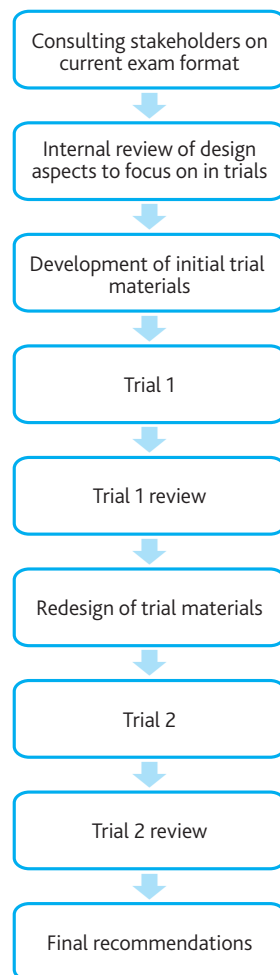


Figure 1: Outline of revision process for Speaking

1. Where reference is made to 'A2 Key', this should be read as inclusive of A2 Key for Schools, and likewise the term 'B1 Preliminary' within this article encompasses the standard and the variants for schools. These exams were previously known as *Cambridge English: Key* and *Cambridge English: Preliminary*.

Consulting stakeholders

As part of its cyclical review process, Cambridge English gathered feedback from over 500 stakeholders (Teachers, Heads of English/English Co-ordinators, Directors of Studies, Centre Exams Managers and Exam Administrators) across several countries (Spain, Italy, Russia, Greece, Romania, Cyprus and Serbia) as well as from a number of assessment experts (Professional Support Leaders, Team Leaders, Chairs, Item Writers), so that both the needs of test users and the expertise of assessment specialists could inform the revision of the Speaking component of A2 Key and B1 Preliminary as effectively as possible. Feedback was collated on the basis of findings from a large-scale survey and focus groups, with key stakeholders including teachers and Centre Exams Managers.

A2 Key

The current test structure is provided in Table 1.

Table 1: Current A2 Key Speaking format

	<i>Interaction pattern</i>	<i>Input</i>
Part 1 Interview 5–6 minutes	Interlocutor asks questions to each candidate in turn, giving factual or personal information.	Spoken questions provided by the interlocutor frame.
Part 2 Information exchange 3–4 minutes	Candidates ask and answer questions about factual, non-personal information.	Spoken and written rubrics. Visual stimuli given in the candidate booklet (see Figure 2 for an example).

Consultation activities, drawing on expert opinion as well as insight from focus groups with key stakeholders, highlighted several areas of potential focus for revision trialling. In the case of A2 Key, there was a concern that the current information gap task (see Figure 2) in Part 2 did not differentiate candidates' levels enough, as evinced by one Centre Manager's comment:

'In my experience, the candidates who have shown themselves to have different levels in speaking in class often get similar results at the Speaking test.'



<p>Candidate A – your answers</p> <div style="border: 1px solid black; padding: 10px; width: 250px; margin: 0 auto;"> <p style="text-align: center;">Skateboarding Competition for anyone 11 – 15 years old</p> <p style="text-align: center;">at Green Park 20 June</p>  <p style="text-align: center;">1st prize New Skateboard visit www.citynews.com for more information</p> </div>	<p>Candidate B – your questions</p> <div style="border: 1px solid black; padding: 10px; width: 250px; margin: 0 auto;"> <p style="text-align: center;">Skateboarding Competition</p>  <ul style="list-style-type: none"> ◆ where ? ◆ for children ? ◆ date ? ◆ website ? ◆ what / win ? </div>
--	---

Figure 2: Current A2 Key Speaking Part 2 sample task

There were also indications that the current Part 2 task was not allowing more able candidates to demonstrate their full speaking performance at this level. Dissatisfaction was also reflected in survey feedback (see Figure 3).

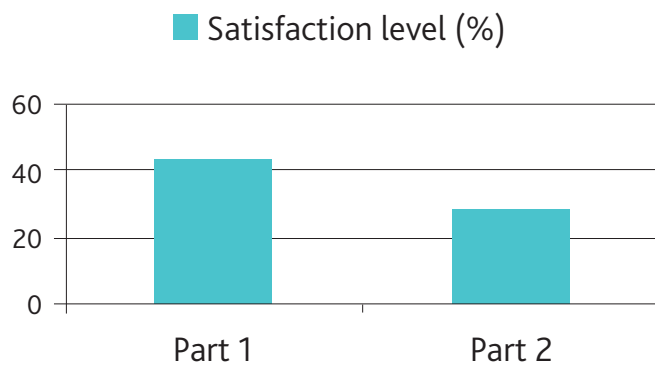


Figure 3: A2 teacher perceptions of the current A2 Key exam format

Survey findings also revealed that developing learners' speaking ability at A2 level was a major concern for teachers. As one teacher commented, their main challenge was 'to encourage students to speak'; another said that 'the greatest challenge is to make them fluent in speaking and taking turns'. Feedback of this kind suggested, at least in part, that the washback effect of the current exam format was not conducive to building learners' interactive skills, for example simple turn taking.

Greater alignment with Level A2 of the CEFR so that candidates have the opportunity to fully demonstrate their ability across a broader range of speaking sub-skills and language functions was thus of fundamental importance in the revision.

According to the CEFR (Council of Europe 2001:58), candidates who have reached A2 level can:

- give a simple description of people, living (and working) conditions and daily routines as a short series of simple phrases and sentences
- explain what they like or dislike about something
- show their ability to manage simple and routine exchanges of ideas and information on familiar topics, provided the other person helps if necessary.

In line with survey feedback, the new test design would aim to provide better measurement and better support for teachers as they prepare their learners for A2 Key Speaking (positive washback).

B1 Preliminary

Table 2: Current B1 Preliminary Speaking format

	<i>Interaction pattern</i>	<i>Input</i>
Part 1 Interview 2–3 minutes	Interlocutor asks candidates questions to elicit personal information.	Spoken questions provided by the interlocutor frame.
Part 2 Collaborative task 2–3 minutes	Interlocutor delegates a collaborative task (simulated situation) to the pair of candidates.	Spoken rubrics. Visual stimuli given in the candidate booklet (line drawings).
Part 3 Long turn 3 minutes	Interlocutor delegates an individual task to each candidate.	Spoken rubrics. Visual stimuli given in the candidate booklet (one photograph per candidate).
Part 4 Discussion 3 minutes	Interlocutor initiates a discussion between the candidates.	Discussion set up by interlocutor using interlocutor frame.

In the case of B1 Preliminary, comments from experts and stakeholders during focus groups tended to concern task order rather than task design. Of particular importance were expert appraisals of how effectively the current Part 3 (collaborative task) and Part 4 (extended discussion) tasks matched their original aims as a consequence of the order in which they appear. Part 4 is intended to be an interactive task following on from the 'long turn' but its reliance on the Part 3 content had the effect of limiting how generative it can be. As one assessment specialist with extensive examining experience noted, Part 3 is always constrained by the photos preceding it; this can result in questions which do not generate very much language, especially for stronger candidates.

There was also a concern about the level of agency assumed by the Part 4 task in that candidates were more or less left to manage the interaction entirely on their own. This was something commented on by one external assessment specialist:

'Currently the aim in Part 4 is for candidates to interact with no support (examiners giving this only if necessary). Part 4 in higher-level tests (B2 First, C1 Advanced and C2 Proficiency) is conducted by the examiner, but allowing for candidates to develop their answers. It could be argued that B1 level candidates actually need more support than the higher levels.'

Comments from several experienced examiners suggested that the final interactive element would more logically follow from the Part 2 discussion task rather than the long turn.

Across both A2 and B1 exams, there was an additional concern to create greater 'family resemblance' across the exam suite as a whole so that from A2 to C2 there is consistency in exam structure as far as is feasible. This enables reduction of test anxiety among learners and also supports teachers preparing students for the exam by standardising the test format.

Having gathered extensive feedback from all relevant sources, assessment specialists focused on developing initial trial test specifications. This information was used to judge how to improve measurement of the construct. For example, in the case of A2 Key, the new task assesses candidates' ability to 'participate in short conversations in routine contexts on topics of interest' (Council of Europe 2018:85).

The revision of assessment scales is typically a much broader and lengthier activity involving all CEFR levels and all assessments aligned to the scales. (Galaczi, French, Hubbard and Green 2011 outline the Cambridge English approach to such work.) However, the existing assessment criteria were considered to accurately represent the A2 and B1 constructs for Speaking and therefore were not changed.

Trial 1: Tasks used

A2 Key

The revised test structure is provided in Table 3.

Table 3: Revised A2 Key Speaking format

	<i>Interaction pattern</i>	<i>Input</i>
Part 1 Interview 3–4 minutes	Interlocutor asks questions to each candidate in turn, giving factual or personal information.	Spoken questions provided by the interlocutor frame.
Part 2 Discussion task 5–6 minutes	Candidates discuss likes, dislikes and give reasons.	Spoken and written rubrics. Visual stimuli given in the candidate booklet (see Figure 4 for an example).

The format for Part 1 remained the same: an interlocutor-led interview. In Phase 1, the interlocutor asks the candidates questions, in turn, using a standardised script to ensure all candidates have the same opportunity to perform. Candidates give basic personal information of a factual nature.

In Phase 2, candidates respond to personal questions, in turn, on two familiar topics such as daily life, school, leisure activities, family, etc. The first two questions require brief answers only. Each candidate is also asked to give an extended response to a prompt connected to the first two questions ('please tell me something about...'). Previously, interlocutors were given the freedom to decide which questions to address to which candidate(s), and questions were chosen at random across a number of topics.

The revised task provides greater coherence as questions relate to just two topics. The more prescriptive frame also supports examiners in their ability to be consistent thereby ensuring equal opportunity for candidates.

The new Part 2 task takes the form of a collaborative discussion. It provides greater opportunities for candidates to fully demonstrate their speaking ability and their interactive communication skills through a more personalised, authentic and meaningful exchange.

In Part 2 Phase 1 candidates are invited to talk together about a topic. They are provided with visual stimuli and asked if they like the activities, places or things depicted and to say why or why not (see Figure 4).

Do you like these different holidays?

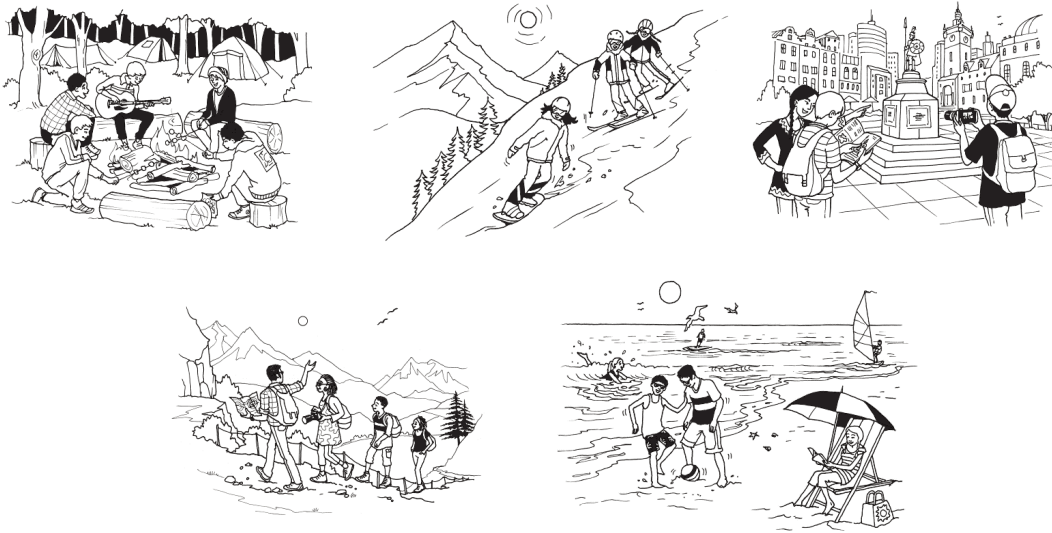


Figure 4: Revised A2 Key Speaking Part 2 sample task

The interlocutor is allowed up to a maximum of two minutes for candidates to talk together independently, before coming in with prompts aimed at extending the discussion and encouraging candidates to develop their utterances, for example, 'Do you think camping is fun?', 'Why (not)?' Interlocutors end this part of the exam with a closing question directed at each candidate in turn; in the case of the sample task: 'Which of these different holidays do you like best?'

Part 2 Phase 1 aims to assess candidates' ability to interact with a partner and with the interlocutor, to express likes and dislikes, and to give simple reasons. Candidates are invited to express opinions about the different activities, things or places represented but are expected to talk about these only in relation to themselves and their experiences of the world, as is appropriate for A2 level.

In Part 2 Phase 2, the interlocutor asks each candidate two more questions, broadening the topic of Phase 1. Phase 2 aims to assess candidates' ability to talk more freely on the topic discussed in Phase 1. Candidates are given the opportunity to demonstrate their full speaking ability in a less formulaic but still supported manner in this last part of the test.

B1 Preliminary

For B1 Preliminary the focus of initial trialling was the following:

- the re-ordering of the tasks, so that the photo-based individual turn task occurs before the discussion task
- the removal of the follow-on question phase from the photo-based individual turn tasks and the inclusion of a follow-up question phase after the Part 3 discussion task (see Figure 5)
- the use of a 'split' rubric in the discussion task (see Figure 6), similar to that of B2 First Speaking Part 3.

Part 4

Interlocutor Use the following questions, as appropriate:

- **What do you do when you want to relax? (Why?)**
- **Do you prefer to relax with friends or alone? (Why?)**
- **Is it important to do exercise in your free time? (Why?/Why not?)**
- **Is it useful to learn new skills in your free time? (Why?/Why not?)**
- **Do you think people spend too much time working/studying these days? (Why?/Why not?)**

Thank you. That is the end of the test.

Select any of the following prompts, as appropriate:

- **How/what about you?**
- **Do you agree?**
- **What do you think?**

Figure 5: Part 4 follow-on questions (revised format)

I'll say that again.

Some students from a small village school are going on a trip to their capital city. Here are some activities they could do there. Talk together about the advantages and disadvantages of doing these activities.

Now, talk to each other.

Candidates

🕒 1-2 minutes

Interlocutor Thank you. Now, I'd like you to decide **which activity would be the most interesting.**

Candidates

🕒 up to 1 minute

Figure 6: Example of split rubric interlocutor frame (Part 3)

Tasks were created to be trialled on both pairs and groups of three.

Trialling cohorts

A2 Key

Eight Speaking Examiners participated in initial trialling, each with at least six years' experience of A2 Key Speaking exams, and, as a group, covering a diverse candidature across Brazil, Czech Republic, Italy, Spain and the UK.

In the trialling, Examiners were invited to watch videos of candidates taking the revised exams and to rate their performances. They were also asked to provide feedback on the new exam by completing a questionnaire.

Candidates were deemed by the Examiners to be typical for the ability level targeted by the exam and to have produced sufficient language to allow Speaking Examiners to rate them across all three assessment criteria.

B1 Preliminary

There was a mix of abilities in this initial trialling cohort but most were students intending to take B1 Preliminary.

The UK-based trialling provided a very diverse range of students from Iran, Albania, Korea, France, Morocco, Saudi Arabia, Japan, Libya, China, Thailand, Colombia, Brazil, Turkey, Czech Republic and Armenia.

Trials overseas featured monolingual pairs, as one might expect, and the sample was intentionally limited in each location in order to provide a representative sample of a diverse range of language groups overall. Despite this, the sample did cover a range of abilities at the level in every case – from strong to weak. Trials were carried out by very experienced Speaking Examiners.

This first phase of trials involved trialling of full tests on over 60 candidates. In all trials, a current-format test was also administered to provide an insight into how the proposed and current designs compared.

Observations and feedback were gathered via standardised forms and following a set of trialling instructions (an excerpt of a trialling observation form is provided in the Appendix). Trials were filmed and in August 2016 a small-scale marking and examiner survey was also carried out using these video recordings.

Following trials, an internal review was conducted where feedback and outcomes were considered.

Trial 1: Review

A2 Key

Overall, Examiners' feedback was positive. The revised tasks were considered to be an improvement on the current format.

Examiners reported that:

- the new tasks provide greater and richer opportunities to assess candidates' interactive communication skills
- the new tasks elicit more real-life realistic language and interaction types than the current ones
- the new tasks are in line with other Cambridge English Qualifications.

Some illustrative comments from the Examiners:

'The previous Part 2 was quite "scripted" and candidates produced a narrower range of language. This task allows the candidates to interact in a meaningful way, and produce language which is their own, rather than relying on the previous prompt questions/information card, which often was misinterpreted or relied on candidates being able to read out loud accurately.'

'It fits better with classroom practice and is less rehearsed.'

'I think the new test format, particularly Part 2, is a better test of language and interactive communication.'

'It's a huge improvement! So much more suitable for this level of candidates and elicits a much wider range of language than the current version.'

Overall, the difficulty level of the new tasks was judged to be appropriate by Examiners, i.e. at A2 level and similar to the difficulty level of the current format. All Examiners agreed that the revised exam allowed stronger candidates to show their speaking ability beyond A2 level. The new format of Part 2 in particular gives candidates greater autonomy, thus enabling them to fully show their ability. The flipside of this was a concern that this could be perceived as a reduction in the support provided to weaker candidates. Indeed, there was a sense in the trialling that the new exam might suit stronger candidates better.

Three out of eight examiners felt that the instructions in Part 2 lacked clarity and that some candidates did not understand the task requirements, leading them to describe the pictures rather than react to them with personal opinions. This lack of clarity could have contributed to concerns around the suitability of the task for weaker candidates.

Finally, feedback from examiners revealed that the anticipated discussion between candidates in Part 2 was not always in evidence as candidates did not respond to each other's utterances. Assessment of candidates' interactive communication was therefore based entirely on their responses to the interlocutor's questions.

Some illustrative comments from the Examiners:

'It wasn't clear that the candidates understood what to talk about with the pictures.'

'Some of them described the pictures in Part 2 rather than giving their opinion of the activities/places, but using the prompts seemed to stop them doing this.'

'...[V]ery few of these trial candidates asked each other a question in the discussion phase – the children in particular just talked about themselves, without really linking what they said to their partner's contribution.'

'When asked to "tell each other" and talk together, some candidates dried up until prompted further.'

These findings fed into recommendations for subsequent trialling. The concept for the new Part 2 task had been proven, but the execution required further refinement; specifically we sought to improve the clarity of the rubric.

B1 Preliminary

The first phase of trials indicated that while the proposed task order was seen as positive based on expert appraisal and evidence from trial footage, the split rubric in the discussion task was problematic in a number of ways for B1 level:

- It often led to repetition, as candidates reached a decision prior to the second rubric being delivered by the interlocutor.
- In the context of this B1 exam, it seemed inauthentic and artificial to divide the appraisal of a range of (relatively concrete) options with the decision of 'which is best', etc.
- At Cambridge English we typically draw a distinction between Speaking at B1 being focused on 'negotiating agreement' and Speaking at B2 stressing 'reaching a decision through negotiation', due in part to the more concrete operations expected of B1 learners.
- From a task-writing perspective, the removal of the focus of the discussion had the effect of making the task scenario inadequate as a springboard for a developed discussion when compared to the current task design. It was hard to 'move the discussion on' at B1.

It was noted in trials that the split rubric was problematic at B1 level:

'Having trialled the split rubric task in isolation from other suggested changes, we were very much of the opinion that the task format wasn't working at this level. The splitting of the task into two phases seems to be artificial. Working towards a conclusion and discussing the relative merits and demerits of the various options is very much one operation in the current test.'

'The removal of the intended outcome of the discussion rendered the context rather thin and made the first part of the task rather abstract. Strong candidates tended to fill this vacuum with an imagined outcome of their own, whilst weaker candidates floundered in a rather abstract discussion that was leading nowhere.'

It could be argued that the demands of a more abstract discussion will generally tend to favour those with greater interactional competence as a result and this was borne out in trials for the split rubric here, where the stronger candidates managed the task better.

In her study of interactional competence, this difference in capabilities is something Galaczi (2014) observes: 'the interactional profile of B1 learners was found to be generally characterised by low mutuality between the speakers' (Galaczi 2014:560). By contrast the B2 learners' better-developed linguistic resources and automaticity in processing allow them to be both focused on constructing their own response and decoding their partner's contributions: 'B2 test takers were found to be more adept at keeping active the roles of speaker and listener at the same time' (Galaczi 2014:564).

Trial 2: Tasks used

A2 Key

The wording of the rubric and the timing allotted in Part 2 Phase 1 were amended in an attempt to provide greater clarity and support, thereby responding to feedback received during initial trialling. A second stage of trialling was subsequently undertaken.

The Part 2 Phase 1 rubric, 'Tell each other what you think about... ' (see Figure 7), was replaced with a direct question – 'Do you like...?' (see Figure 8). By reducing the structural complexity of the instruction, the processing load was lowered and task requirements, it was anticipated, would be clearer and easier to grasp.

Tell each other what you think about these different holidays.
I'll say that again.
Here are some pictures that show different holidays.
Tell each other what you think about these different holidays.
OK? Talk together.

Figure 7: Trial 1 Part 2 Phase 1 rubric

Do you like these different holidays? Say why or why not. I'll say that again.
 Do you like these different holidays? Say why or why not.
 All right? Now, talk together.

Figure 8: Trial 2 Part 2 Phase 1 rubric

To ensure adequate support for weaker candidates, the timing requirements of Part 2 Phase 1 were also revised. In Trial 1, interlocutors were instructed to 'allow up to two minutes for discussion'. In Trial 2 this was reduced to 'allow up to one minute', thus allowing the Examiner to intervene and lend support to candidates earlier.

This was later revised to 'a minimum of one minute and a maximum of two' to allow Examiners the flexibility to tailor timing requirements to the needs of candidates. While some candidates at this level could only sustain interaction without help from the interlocutor for one minute, some candidates appeared able to sustain it for longer. In the case of stronger candidates, or candidates who took time to warm up, interjecting after one minute risked interrupting the flow and not giving candidates the opportunity to extend their discussions fully.

B1 Preliminary

A summary of the revised test structure is provided in Table 4.

Table 4: Revised B1 Preliminary Speaking format

	<i>Interaction pattern</i>	<i>Input</i>
Part 1 Interview 2–3 minutes	Interlocutor asks candidates questions to elicit personal information.	Spoken questions provided by the interlocutor frame.
Part 2 Long turn 3 minutes	Interlocutor delegates an individual task to each candidate.	Spoken rubrics. Visual stimuli given in the candidate booklet (one photograph per candidate).
Part 3 Collaborative task 2–3 minutes	Interlocutor delegates a collaborative task (simulated situation) to the pair of candidates.	Spoken rubrics. Visual stimuli given in the candidate booklet (line drawings).
Part 4 Discussion 3 minutes	Interlocutor leads a discussion with candidates.	Spoken questions provided by the interlocutor frame.

The 'split' rubric was not included in the second stage of trialling for the reasons outlined in the section on Trial 1. Instead, a version of the current task rubric was developed which avoided the repetition of the rubric (see Figure 9).


Part 3	
Interlocutor	Now, in this part of the test you're going to talk about something together for about two minutes. I'm going to describe a situation to you. <i>Place Part 3 booklet, open at Task 1, in front of the candidates.</i> A young man works very hard, and has only one free day a week. He wants to find an activity to help him relax. Here are some activities that could help him relax. Talk together about the different activities he could do, and say which would be most relaxing. All right? Now, talk together.
Candidates
 <i>approx. 2–3 minutes</i>	
Interlocutor	Thank you. (Can I have the booklet please?) <i>Retrieve Part 3 booklet.</i>

Figure 9: Example of Phase 2 Part 3 rubric (non-split)

Another significant change in Phase 2 was the uncoupling of the photos used in the individual turn tasks. Previously these were linked thematically (e.g. 'A day out') but this was felt to potentially advantage or disadvantage candidates, while also making the successful development of these tasks far more difficult.

For Phase 2, the photographs used were deliberately paired so that they would *not* overlap in basic topics or themes. There were also back-up prompts (see Figure 10) added to the interlocutor frame for the individual turn to provide additional means for interlocutor support.

Back-up prompts
<ul style="list-style-type: none"> • Talk about the people/person. • Talk about the place. • Talk about other things in the photograph.

Figure 10: Back-up prompting in Part 2 (individual turn)

There were also more minor alterations to Part 1, based on observations from the first phase of trials (e.g. slight changes to the introductory rubrics to ensure a better flow of questions).

Trialling cohorts

A2 Key

Seven Senior Examiners administered the revised tasks to 127 candidates across six countries (Argentina, Greece, Romania, Russia, Taiwan and the UK). Qualitative analysis of their feedback was conducted. Seventeen Russian candidates sat the pre-revision and the revised Part 2 tasks so that a direct comparison of candidate performance across the two formats could be made. This was completed via functional analysis of candidates' speech, achieved by comparing the number of language functions elicited by both formats (see Figure 11).

Candidates were deemed by the Examiners to be representative of the target candidature for the test and to have produced sufficient language to allow Speaking Examiners to rate them across all three assessment criteria.

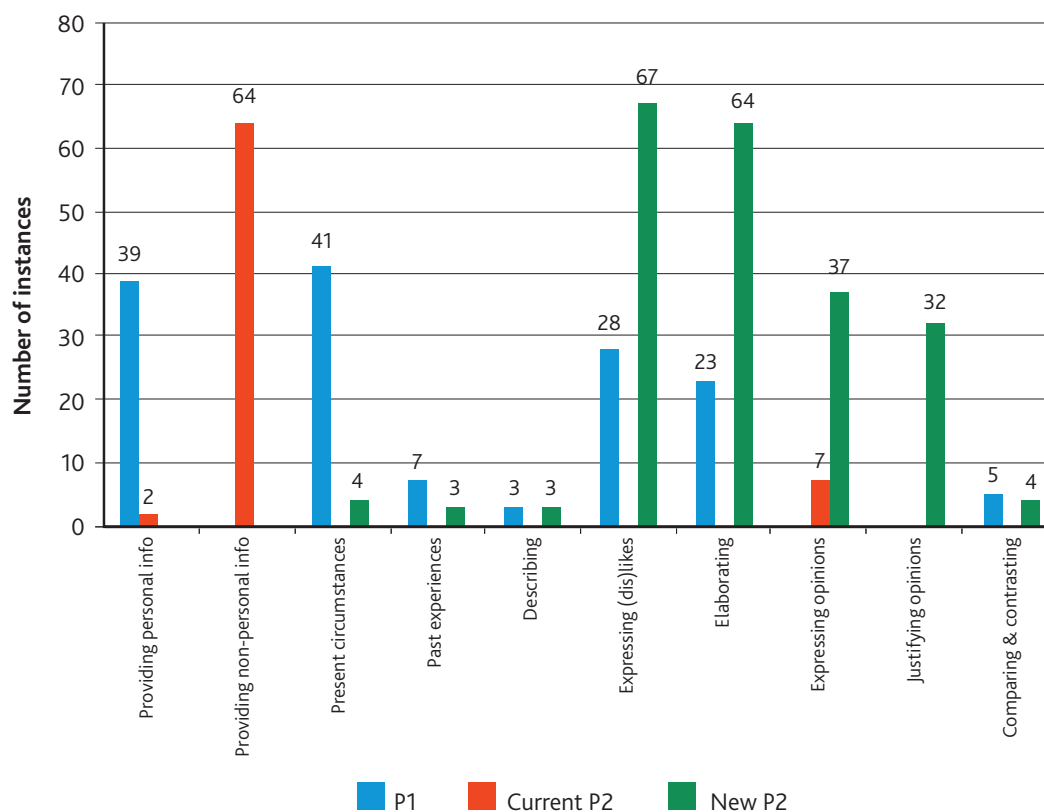


Figure 11: Functional analysis of language in trialling of A2 Key Speaking (P = Part)

B1 Preliminary

Phase 2 trials were carried out in a similar wide range of locations and with an equally wide range of ability levels as in Phase 1. As in the first phase, trial materials were complete tests and these were carried out alongside a current-format test to provide comparison data and feedback. More than 80 candidates were involved in this phase in locations in Russia, Argentina, Taiwan, the UK, Italy, Romania, Vietnam and Greece. All samples were recorded for analysis, examiner surveys and feedback activities.

Trial 2: Review

A2 Key

Examiners were consistent in their opinions that the revised exam, and especially Part 2, was an improvement on the pre-revised task. Themes that stood out from their feedback are as follows:

- it allows candidates greater flexibility and more opportunities to demonstrate their A2-level knowledge and speaking skills, particularly their interactive skills, more fully
- it allows especially strong candidates to demonstrate their speaking skills beyond the requirements of A2 level
- it elicits more personalised language: candidates can produce novel utterances
- it elicits a wider range of language functions, e.g. responding to opinions and turn-taking

- it focuses more on meaning than on form, thereby indicating its potential for positive washback, i.e. a focus on communicative language use rather than formal accuracy
- it allows for more meaningful and more authentic interaction
- it increases candidate enjoyment and overall performance
- it is in line with other Cambridge English Qualifications.

Some illustrative comments from the Examiners:

'I totally like this new format and I think it can give candidates better opportunities for speaking and using more language. Pictures give candidates more independence.'

'The difference [in the new Part 2 versus current] in the quality of students' utterances and their resultant performances was stark ... students consistently performed so much better in the revised Part 2.'

'Students responded very well ... and participated in lively discussions. Even though they were not prepared for a task like this, they managed to sustain a simple discussion.'

'One of the students said: "I love to talk about things with my friends. The other thing we did was not so interesting.'"

B1 Preliminary

Extensive feedback was taken from Examiners involved in the second phase of trials. The vast majority of this feedback endorsed the new proposed test design, which was felt to 'flow' more naturally and allow the candidate time to warm up via the individual turn prior to the collaborative discussion task.

No evidence from trialling suggested that the use of different topics/themes in the photo-based task would disadvantage either candidate. The use of different topics/themes also limits the potential for candidates to 'lift' language from each other.

In Part 3, the revised discussion task rubric indicated in trials that it worked well and that the removal of the repetition of the main rubric was not impacting on the candidates' ability to perform the task.

The use of a follow-on set of questions in Part 4, after the discussion task, was also felt to be a positive move as it meant B1 candidates were no longer required to take on an interlocutor-like role in the interaction and the examiner was more able to re-balance the contributions from candidates, as in B2 First, B2 First for Schools and C1 Advanced Part 4.

This was seen as preferable to the current B1 Preliminary Part 4 task, which often elicited two further 'long turns' from candidates rather than a genuine interaction. This also meant the examiner could step in if candidates 'dry up' in their response while also giving scope for some further interaction (i.e. by the interlocutor directing a question to both candidates). The new Part 4 still afforded the assessors scope to fine-tune their marks in the final phase of the test event too.

Conclusions

A2 Key

In conclusion, the new exam is considered to be an improvement on the current format. It elicits a wider range of language and language functions, thus allowing candidates to fully demonstrate their speaking skills and providing a more authentic and meaningful task. Despite the resounding positive appraisal of the new tasks, Cambridge English remains mindful of the need for clarity of instruction and expectations for those who sit the exams as well as for those who administer them. As part of the rollout of the new format, we will ensure that we:

- provide information on the focus of Part 2 to candidates and their teachers, highlighting that the pictures are intended to prompt discussions about the activities, places and things represented and that the task is not to describe them, and that candidates should be encouraged to respond to their partner's utterances
- include advice and appropriate back-up questions for Speaking Examiners to help guide candidates through the task, to provide appropriate scaffolding and support, and to allow candidates to demonstrate their speaking skills fully.

B1 Preliminary

In conclusion, it was felt the revised exam format for B1 Preliminary was one that provided much greater interlocutor control than the existing test design, and improved the test experience for candidates without diminishing the test's ability to make accurate assessments of candidates at this level. The focus on the CEFR B1 level is maintained, but the revised test also allows stronger candidates to show a fuller range of skills and also aims to support less able candidates more than previously.

References

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

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

Galaczi, E (2014) Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics* 35 (5), 553–574.

Galaczi, E, French, A, Hubbard, C and Green, A (2011) Developing assessment scales for large-scale speaking tests: A multiple-method approach, *Assessment in Education: Principles, Policy & Practice* 18 (3), 217–237.

Appendix: Excerpt from observation form used in trialling

For quantitative observations of language functions, the following type of form was used:

Informational functions/features	Part 1	Part 2	Part 3	Part 4
Providing personal information				
Providing non-personal information (e.g. dates/times/prices)				
Talking about present circumstances				
Talking about past experiences				
Talking about future plans				
Describing people, places and situations				
Expressing preferences				
Expressing opinions				
Justifying opinions				
Elaborating (e.g. explaining and giving reasons)				
Comparing and contrasting				
Suggesting and recommending				
Any other functions? [Please state below]				

Informational functions/features	Part 1	Part 2	Part 3	Part 4
Agreeing/disagreeing				
Asking for opinion				
Asking for information				
Negotiating meaning: - checking meaning - asking for clarification - correcting utterance(s)				
Responding to required clarification				
Paraphrasing words and ideas if necessary				
Any other functions? [Please state below]				

Informational functions/features	Part 1	Part 2	Part 3	Part 4
Responding appropriately				
Initiating				
Turn-taking				
Sustaining a long turn				
Any other functions? [Please state below]				

The following examples of trial feedback questions are taken from a standardised form focused on more qualitative judgements on trialling of the 'split' rubric in B1 Preliminary Speaking (Phase 1):

How would you describe the candidates' response to the 'split rubric'?

For example:

- Was the transition from one phase of the task to the other smooth?
- Did any of the rubrics require repetition or clarification?
- As an Examiner, did you feel the rubrics were easy to read/manage?
- Were the rubrics clearly understood?
- Was there overlap or repetition in language produced between the discussion and decision phases?
- How did the suggested timing for the candidate response compare to reality?

How would you describe the candidates' response to the Part 4 task?

For example:

- As an Examiner, did you feel the rubrics were easier to read/manage?
- In comparison with the current *Cambridge English: Preliminary* and *Cambridge English: Preliminary for Schools* Part 4 task, did you feel this was better in terms of managing the test experience for candidates?
- In comparison with the current *Cambridge English: Preliminary* and *Cambridge English: Preliminary for Schools* Part 4 task, did you feel this was better in terms of providing a good sample of language (particularly re: 'fine-tuning' of marks)?
- In comparison with the current *Cambridge English: Preliminary* and *Cambridge English: Preliminary for Schools* Part 4 task, did you feel this was better in terms of providing an equal contribution from both candidates?
- How did the suggested timing compare with the real time taken?
- Were all the questions you used clear and understood?
- How many questions did you use in total? Did you make much use of the additional prompts (e.g. 'What do you think?')?

Revising the A2 Key and B1 Preliminary Writing exam

Anthi Panagiotopoulou Assessment, Cambridge Assessment English

James Lambie Assessment, Cambridge Assessment English

Kevin Y F Cheung Research and Thought Leadership, Cambridge Assessment English

Introduction

The revised A2 Key, A2 Key for Schools, B1 Preliminary and B1 Preliminary for Schools will be introduced in 2020.¹ The major drivers behind changes to the assessment of writing for these Cambridge English Qualifications were to ensure the Writing test constructs remain fit for purpose, and to better support upward certification, which was introduced in 2011.

In the pre-revision Key exam, writing was assessed as part of the Reading and Writing papers, and performance for these two skills was reported together. Candidates were asked to produce little of their own writing, both in Key and Key for Schools. The most noticeable change to how writing will be assessed post-revision is the reporting of Writing scores separately from scores for reading. For Preliminary, Reading and Writing will be separated into two distinct papers. Key will also report Reading and Writing scores separately, but will continue to combine the tasks in a single paper for practical purposes.

This approach aligns Key and Preliminary more closely with other Cambridge English Qualifications, such as B2 First (Lim 2015), increases the proportion of testing that focuses on writing and provides more information about a candidate's writing performance. The latter outcome is particularly important because providing useful information for learners is at the heart of the renewed Cambridge English focus on Learning Oriented Assessment (LOA). Key and Preliminary are potentially 'forms of large-scale evidence which could contribute evidence of learning during rather than at the end of the process' (Jones and Saville 2016:79). Candidates taking Key are often near the beginning of their learning journeys and these revisions are an opportunity to put LOA principles into practice with a large-scale exam. By increasing the proportion of testing time spent on writing, we also intended to create positive washback on the teaching and learning of writing skills at these levels.

In preparation for the revision, a series of structural equation modelling (SEM) studies were conducted with exam data from Key and Preliminary to inform the revision process and explore the constructs assessed across the three papers (Listening, Speaking, Reading and Writing). SEM is a measurement technique used to investigate the underlying constructs assessed by an exam and the relationship between skills targeted by different components. These analyses showed that some tasks targeting writing ability in the pre-revision exams, particularly those in Key, were more closely associated with reading than performance on productive writing tasks. For the pre-revision exams, this had no impact on results because Reading and Writing scores were reported together. In order to report these scores separately, as planned for the exams post-revision, we needed to separate the tasks and ensure that those contributing to the Writing score covered the writing ability construct more adequately. In addition, a panel of experts in writing assessment reviewed the writing papers to establish how well tasks were eliciting writing at the target levels. Therefore, some tasks have been discontinued and others have been specified as assessments of reading ability. To ensure that the writing construct was fully represented, new tasks were designed and trialled; one was added to Key and two new tasks are included in Preliminary.

1. A2 Key and B1 Preliminary will hereafter be referred to as Key and Preliminary, respectively; A2 Key for Schools and B1 Preliminary for Schools will hereafter be referred to as Key for Schools and Preliminary for Schools, respectively.

Key: What has changed and why?

The pre-revision Key and Key for Schools included a guided writing task (Part 9), which asked candidates to write a short message of 25 words or more, as a note or an email. Cambridge English assessment specialists reviewed versions of this task and candidate responses, with reference to the Common European Framework of Reference for Languages (CEFR, Council of Europe 2001). By reviewing responses from live sessions alongside A2 CEFR descriptors, it was established that the task was eliciting responses at the target CEFR level. In particular, candidate submissions were aligned to the following Can Do statements, which are an important part of the writing construct as conceptualised in the CEFR. Therefore, this guided writing task was retained for the exam.

Can write short, simple formulaic notes relating to matters in areas of immediate need.

Can write short, simple notes, emails and text messages (e.g. to send or reply to an invitation, to confirm or change an arrangement).

However, this was the only productive writing task in the pre-revision Key exam and it exclusively focuses on CEFR illustrative descriptors from the writing correspondence scale. The task was also reviewed by the panel of experts alongside ALTE 'Can Do' statements for A1, A2 and B1. This showed that the guided writing task focused on the following A2 statements:

Can write simple letters of the "Thank you" type.

Can convey personal information of a routine nature to, for example, a penfriend.

Can express opinions of the "I don't like" type.

The panel of experts in writing assessment noted that the task did not cover B1 statements for writing and recommended that a new productive writing task be considered as part of the revisions. Assessment specialists responsible for setting pre-revision papers also recommended trying to elicit more writing. For example, when asked to provide feedback, they reported that:

'Continuous writing is not given enough importance. I think there could be a place for a longer writing task that carried more of the overall marks.'

'Replace [Part 8] with an additional writing task, e.g. story or simple essay to stretch the more able candidates.'

'The current [Key] Part 9 is very similar to Part 2 on the current PET [former name for Preliminary] Writing paper. The number of words that candidates have to produce for Part 9 is, at 25–35 words, very low. Since there should be scope for candidates to demonstrate ability at B1 level on the [Key] paper it may be worth considering including another writing task.'

In addition to increasing the variety and amount of writing required of candidates, the panel recommended that a new task should better support upward certification to B1 level.

The other parts specified as writing in the pre-revision Key exam (6–8) were a spelling task, an open cloze task and information transfer tasks. As previously mentioned, SEM analyses indicated that these tasks were assessing skills more associated with reading than with writing. Other analyses also identified further issues to consider as part of the revisions. For example, we investigated how similar-ability candidates from different groups perform on particular tasks. This analysis, known as differential item functioning (DIF), indicated that some groups (particular L1s or age groups) performed differently on the spelling task. Some of these findings may result from different teaching practices across geographical regions, particularly where some countries focus on particular topic areas

for learning vocabulary. Although this is not problematic, and there are other possible explanations for this finding, the panel of experts in writing assessment recommended reducing these group differences in the revised exam, if possible.

Teachers and candidates were also surveyed to gather their perspectives on the paper and the tasks; they identified the information transfer task as unpopular, because it was difficult to match to commonly used curricula. The spelling and information transfer tasks were removed from the exam based on the reviews by assessment specialists and the research studies. Also informed by the SEM studies, the open cloze task was re-specified as an assessment of reading ability for inclusion in the revised Key exam. These changes made sufficient space for us to consider another productive writing task.

According to the CEFR, A2-level language learners should be able to write simple connected texts, so new tasks to elicit extended writing were proposed, designed and trialled. One of these tasks was selected for inclusion in the revised exam. The new 'picture story' task requires test takers to write a short 35-word story based on three pictures. This allows the candidate to demonstrate they can perform functions from the creative writing illustrative descriptors of the CEFR (see Table 1), complementing the aspects of writing targeted by the guided writing task.

Table 1: Creative writing descriptors from CEFR A2 and B1 (Council of Europe 2018:76)

B1	Can narrate a story. Can clearly signal chronological sequence in narrative text.
A2	Can tell a simple story (e.g. about events on a holiday or about life in the distant future). Can write very short, basic descriptions of events, past activities and personal experiences.

In most writing tasks at this level (including the guided writing task retained as Part 6), content points are provided in bullet points to scaffold test takers and support them to produce text. This design feature reduces the cognitive load of generating ideas, freeing up resources needed to access lexical and syntactic knowledge (Shaw and Weir 2007). However, the bullet points can make it difficult to assess organisation at lower levels, because test takers inevitably match the presentation order of the bullet points precisely in their written response. An alternative approach to providing content points was adopted for the new writing task, to give test takers the opportunity to construct their own organisational structure. Instead of written bullet points, pictures are provided to help test takers generate ideas. These pictures give test takers something concrete to base their responses on, but they also allow for more creativity to be used in generating ideas and structuring responses. Development of this task was informed by recent work on the assessment of writing in the Cambridge English Young Learners exams (Davies and Dunlop 2018). Following a series of trials, the new task was specified and developed for inclusion in the Writing component. The result is a new Key Writing test that includes two productive writing tasks, covering key functions of writing included in the CEFR (Council of Europe 2018). These additions, and previously discussed changes, meant that there were some alterations to the overall format of the Key Writing paper (see Table 2).

The new picture story task encourages test takers to engage in some simple organising of their responses. Instead of only testing candidates' ability to produce simple transactional messages, we are now testing their ability to produce simple narratives. The task and the accompanying mark scheme enable us to distinguish between candidates who 'can write very short, basic descriptions of events, past activities and personal experiences' (A2) and those who 'can narrate a story' (B1). Trials of the revised Key Writing paper showed the new format successfully achieved this aim.

Table 2: Changes in the Writing paper of the Key and Key for Schools Reading and Writing papers

	<i>Task type</i>	<i>What is required of the candidate</i>	<i>Rationale for decision</i>
Removed	Part 6: Spelling task	Read five definitions and write the words they define – the first letter and number of letters are given.	1. SEM studies indicated that the task did not cover sufficiently the sub-construct of writing ability. 2. DIF studies showed some unexpected group differences in task performance.
	Part 8: Information transfer task	Read two short input texts and complete gaps in output text.	1. SEM studies indicated that the task did not cover sufficiently the sub-construct of writing ability. 2. Survey studies showed that this task was unpopular with teachers and candidates, due to being difficult to match with curricula.
Revised for the Reading paper	Part 7: Open cloze	Completing gaps in text.	SEM studies showed that this task was assessing skills more associated with reading than with writing.
Retained	Part 9: Guided writing (Part 6 in revised Key)	Write a short email or note of 25 words or more.	Analysis of candidate responses by expert reviewers indicated that this task was appropriate for eliciting A2 writing.
Introduced	Picture story (Part 7 in revised Key)	Write a short story of 35 words or more based on three picture prompts.	Introduces a longer productive writing task to Key, allowing candidates to display a wider range of writing skills that cover some CEFR B1 descriptors.

Preliminary: What has changed and why?

Preliminary reading and writing tasks were administered in a single paper pre-revision even though the two skills were reported separately. The Writing paper consisted of a sentence transformation task, a short communicative message task, and an extensive writing task, where candidates were able to choose between writing a story or an informal letter.

In order to explore whether the exam remains fit for purpose, a construct investigation study was carried out to examine whether the underlying constructs of the exam support a componential aspect of language proficiency where each component assesses a distinct aspect of language proficiency (Elliott, Docherty and Benjamin 2015). Exploratory Factor Analysis (EFA) was carried out on the papers to map the task types to the skills measured and then Confirmatory Factor Analysis (CFA) was used to investigate the appropriacy of different plausible construct models.

The analyses showed that the Writing paper would benefit from being revised as the sentence transformation task was tapping more into the reading construct than into the writing one and the short communicative message task did not clearly measure writing ability. The extensive writing task was found to clearly represent the intended writing construct and the use of assessment criteria allowed for accurate score interpretation. For this reason, it was decided to replace the sentence transformation task and the short communicative message task with one or more extensive writing tasks that would represent the writing construct at B1 more clearly, and which would be assessed using the same assessment criteria as the short communicative message task, thus increasing both the validity and the reliability of the paper.

The email task in Preliminary and Preliminary for Schools is one of the new tasks and has replaced the pre-revision optional informal letter task. It requires the candidate to demonstrate the ability to handle the language of functions, retaining the testing focus of the current short email message but at the same time eliciting a fuller range of functions expected at B1. The candidate is supported through the provision of annotations on the email, which provide the necessary scaffolding for the level. The required output being at 100 words, though, enables the candidate to display a wider range of writing sub-skills, thus allowing for more accurate assessment of candidates' performance and upward certification. Table 3 gives a full summary of the task changes.

Table 3: Changes in the Writing paper of Preliminary and Preliminary for Schools

	<i>Task type</i>	<i>What is required of the candidate</i>	<i>Rationale for decision</i>
Removed	Sentence transformations	To linguistically manipulate part of a sentence to make it semantically equivalent to an input sentence.	Often constrained to one-word answers, which are not part of the B1 writing construct.
	Short communicative message	To produce a short communicative message based on three prompts.	Second longer writing task more suited to level and upward certification.
	Informal letter (optional)	To produce an informal letter based on a written prompt.	Too similar in style to new compulsory email, so retesting the same part of the writing construct.
Retained	Story writing (optional)	To produce a story using a given title or first sentence.	
Introduced	Email response to annotated input (compulsory) Change from 30 words to 100	To respond to an email based on annotations on the input email.	Retains functional focus of current short message, but allows stronger candidates to show B2 level.
	Article (optional)	To produce an article on a topic of personal interest, describing events, feelings and giving opinions.	Allows candidates to display a wider range of writing skills, improving validity of upward certification.

The email genre was chosen to be the compulsory task in Preliminary on the basis that non-native speakers are most likely to have to cope with emails in English in their future lives and most candidates already have genre knowledge, i.e. knowledge of the genre conventions which help candidates shape possible responses (Matsuda and Silva 2010). Having the email task as a compulsory one provides a basis for comparison between candidates and brings the structure of the test in line with the structure of B2 First and B2 First for Schools.

Part 2 of the Writing paper in Preliminary and Preliminary for Schools is an extensive writing task which is not supported by annotations as Part 1 is. Candidates are required to draw on their linguistic resources in order to respond to either an article task, which is new to the paper, or a story task, which is present in the current version of the paper. At this level of writing, candidates are asked to create pieces of extensive writing that demonstrate coherence. This enhances the performance authenticity (Brown and Abeywickrama 2010) and widens the range of lexical resources the candidate can use. This will better enable candidates whose writing ability is above CEFR Level B1 to demonstrate it and be awarded marks accordingly.

Having the email task as a compulsory Part 1 meant that a new task was needed as an option for Part 2 of the paper. Several possible genres were considered, but it was found that the article task was a better match to the Preliminary candidates' experiences and CEFR B1 creative writing descriptors. The candidates are provided with prompts to

guide their writing and they are asked to draw on their own experiences in order to write the article addressing the prompts. They are asked not only to use descriptive language, but also to write accounts of their feelings, reactions and opinions, thus broadening the construct. Creative writing is an important part of the writing construct at B1 level which can allow candidates to show their linguistic range and control.

The story task has remained as an option for Part 2 as the construct investigation analyses (Elliott et al 2015) showed that this specific task not only represented the construct well but also allowed for valid inferences to be made about the candidates' proficiency level.

As Table 4 shows, CEFR descriptors for creative writing apply to both tasks of the Writing paper Part 2 and would thus contribute to the generalisability of reported scores and reliability of the exam, regardless of which task a candidate would opt for.

Table 4: CEFR creative writing descriptors (Council of Europe 2001:62)

B2	<p>Can write clear, detailed descriptions of real or imaginary events and experiences, marking the relationship between ideas in clear connected text, and following established conventions of the genre concerned.</p> <p>Can write clear, detailed descriptions on a variety of subjects related to his/her field of interest.</p> <p>Can write straightforward, detailed descriptions on a range of familiar subjects within his/her field of interest.</p>
B1	<p>Can write accounts of experiences, describing feelings and reactions in simple connected text.</p> <p>Can write a description of an event, a recent trip – real or imagined.</p> <p>Can narrate a story.</p>

Reliability is further increased by the fact that all tasks in the Writing paper are now marked using the same assessment scales that were themselves developed with reference to the CEFR. These scales were in place pre-revision and were used to mark the optional tasks (informal letter and story) and were found to serve the construct appropriately.

Including a compulsory extensive writing task meant that the candidate output as well as the time required for the test would be increased. Taking into consideration the exam's cognitive load as well as the average age of the candidature, the panel recommended that it would be for the candidates' benefit to separate the Reading and Writing paper. The revised Writing paper includes two 100-word tasks for the candidates to complete in 45 minutes and is administered after the Reading paper has been completed. This allows for the candidates to fully focus on the Writing paper and manage their time efficiently, and timing was explored in trials of the test.

The changes at both Key and Preliminary also mean that writing – defined as the production of written text for a communicative purpose (as opposed to notes or single words) – is receiving greater prominence among the four skills. As a result, Writing scores for Key and Key for Schools can now be reported separately from the Reading scores, and on Preliminary and Preliminary for Schools, Writing is now a separate paper. Table 5 summarises how the revision process has changed the Writing component.

Table 5: Increased prominence of productive writing

		<i>No. of parts involving the production of written text</i>	<i>No. of words candidates should write</i>	<i>Recommended time candidates will spend writing</i>	<i>Marks available for communicative writing</i>	<i>Reporting of the Writing marks</i>
Key and Key for Schools	Pre-revision	One (out of nine on the Reading and Writing paper)	25–35	10 mins out of 70	5 (out of 60 for the Reading and Writing paper)	Reading and Writing reported together
	2020 revision	Two (out of seven on the Reading and Writing paper)	60–80	20 mins out of 60	30	Reading and Writing reported separately
Preliminary and Preliminary for Schools	Pre-revision	Two (out of eight on the Reading and Writing paper)	135–145	40 mins out of 90	15 (out of 25 for the Writing paper)	Reading and Writing reported separately
	2020 revision	Two (but Writing is now a separate paper from Reading)	About 200	A separate paper taking 45 mins	40	Reading and Writing reported separately

How were the changes received?

We have seen that the revised papers are intended to better reflect the writing constructs for their level and to increase the prominence of writing at both levels, but now we need to look at whether candidates and teachers agree. Trialling showed that the new Key and Key for Schools picture story task is perceived by candidates as one of the more difficult parts of the paper, while still being manageable for most candidates, and is perceived as more difficult than the email task. The task type of each part of the test is given in Table 6 and Figure 1 breaks down how each part was rated for difficulty by candidates.

Table 6: Format of revised Key and Key for Schools exam

Part 1	3-option multiple choice with messages and signs
Part 2	Multiple-matching reading
Part 3	3-option multiple choice reading (long text)
Part 4	Multiple-choice cloze
Part 5	Open cloze
Part 6	Guided email writing
Part 7	Picture story

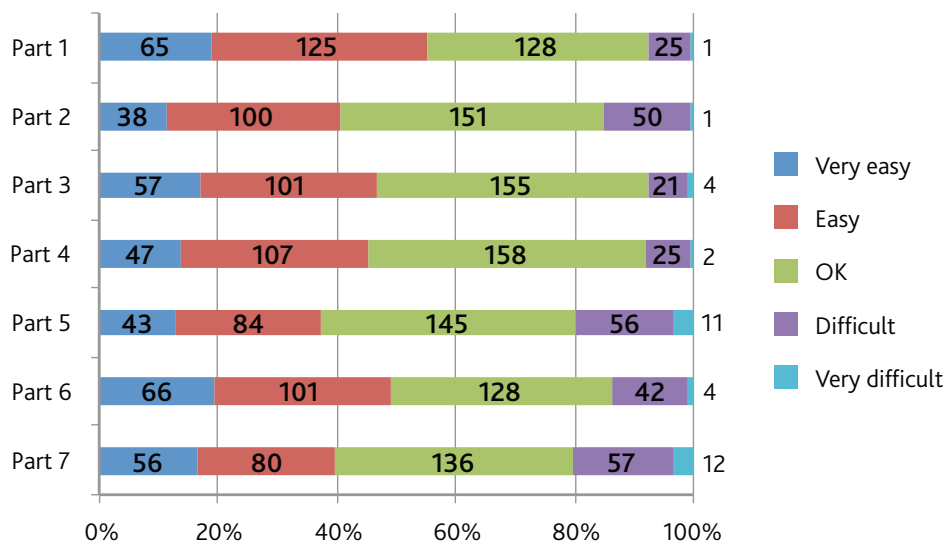


Figure 1: Revised Key and Key for Schools tasks rated for difficulty by candidates

Comments from teachers also suggested that the picture story made the paper more difficult, but again they did not suggest this was inappropriately difficult.

'The writing is longer, however I think it could benefit students.'

'Writing a story implies better handling of tenses and cohesive devices.'

'[What I like about this test is] story telling – it makes the exam more challenging.'

Similarly for B1 Preliminary and B1 Preliminary for Schools, 95% of candidates who took the trial tests agreed that the test allowed them to show their writing ability and that task instructions were clear (see Figure 2). Candidates, teachers, invigilators, stakeholders and focus groups provided feedback on paper difficulty, writing amounts and timing through a variety of channels. Candidates reported that the level of the new tasks was deemed appropriate, the amount of writing was neither too little nor too much, and the proposed timing of the paper at 45 minutes was considered sufficient. Other test users reported that 'the tasks were important and worth students' time to prepare'. Test users also acknowledged the improvements made to the test and reported that the revised paper allowed candidates to better show their language ability (Vidaković and Elliott 2018).

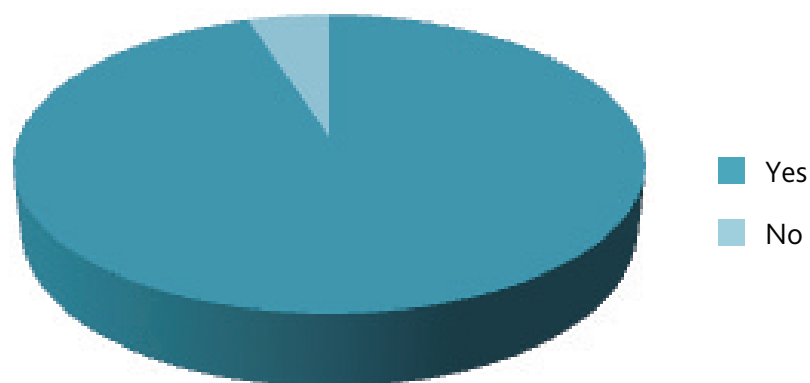


Figure 2: Candidate feedback, Trial 2 – The test allowed me to show my writing ability

As with Key and Key for Schools they also found the second writing task more difficult – 12% responses at 'difficult' or 'very difficult' for Part 1, but 21% responses at 'difficult' or 'very difficult' for Part 2.

What is the expected impact on teaching and learning?

We expect that the changes will have a positive effect on teaching and learning. With productive writing now asking candidates to do more independent writing on Key and Key for Schools, and being made a separate paper on Preliminary and Preliminary for Schools, we expect it to feature more prominently than before in teachers' course and lesson planning. Learners will be required to produce more texts (emails, stories, articles) and will receive more feedback on the aspects of their writing that are tested (lexis, grammar, organisation, task achievement). This should encourage teachers to devote more class time to productive writing, and particularly to practising extended writing. In contrast to short writing exercises, which typically focus on lexis and grammar, extended writing exercises provide important opportunities to raise awareness of other aspects, such as cohesion.

Research is currently underway to establish baseline data for A2 Key for Schools. We will be interviewing teachers to find out how much classroom and homework time is devoted to productive writing, and how they focus on it. Once the revision is introduced we will go back to ascertain what effect the new picture story task and the overall increased prominence of productive writing have on the test as a whole, have had on teachers' approach to course and lesson planning, and to classroom practice.

Conclusion

The revision of Key, Key for Schools, Preliminary and Preliminary for Schools resulted in important changes to the way writing skills are assessed. The key consideration at both A2 and B1 levels was to test more of the writing construct. The driver behind this was ensuring that the test constructs remain fit for purpose, especially in relation to upward certification. This has been achieved through the inclusion of new tasks that target CEFR descriptors which were previously not tested. As a result, candidates are being asked to do more independent writing than on the pre-revision tests, which allows for the reporting of separate scores for Reading and for Writing on Key and Key for Schools, and for Writing to become a separate paper on Preliminary and Preliminary for Schools. This, along with the new task types, brings both papers more into line with other papers from B2 First upwards. It is anticipated that the increased prominence of writing at A2 and B1 will be reflected in a proportionate increase in the amount of study time devoted to the skill, which will provide candidates with a good grounding as they progress towards B2 First.

References

- Brown, H D and Abeywickrama, P (2010) *Language Assessment: Principles and Classroom Practices*, White Plains: Pearson Education.
- Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Cambridge: Cambridge University Press.
- Council of Europe (2018) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors*, Strasbourg: Council of Europe.
- Davies, K and Dunlop, M (2018) New writing tasks for young learners, in Cambridge Assessment English (Ed) *The 2018 Pre A1 Starters, A1 Movers and A2 Flyers Revisions*, Cambridge: Cambridge Assessment English, 25–27, available online: www.cambridgeenglish.org/Images/461823-young-learners-revision-publication.pdf
- Elliott, M, Docherty, C and Benjamin, T (2015) *PET(fs) Construct Investigation (SEM)*, Cambridge: Cambridge English internal report.
- Jones, N and Saville, N (2016) *Learning-oriented Assessment: A Systemic Approach*, Studies in Language Testing volume 45, Cambridge: UCLES/Cambridge University Press.
- Lim, G S (2015) Aspects of the revision process for tests of Writing, *Research Notes* 62, 32–35.
- Matsuda, P K and Silva, T (2010) Writing, in Schmitt, N (Ed) *An Introduction to Applied Linguistics*, London: Routledge, 232–246.
- Shaw, S D and Weir, C J (2007) *Examining Writing: Research and Practice in Assessing Second Language Writing*, Studies in Language Testing volume 26, Cambridge: UCLES/Cambridge University Press.
- Vidaković, I and Elliott, M (2018) *Preliminary (fs) Revision Sep 2016 – Across skills*, Cambridge: Cambridge English internal report.

We help people learn English and prove their skills to the world

Discover more:
cambridgeenglish.org/researchnotes

We are Cambridge Assessment English. Part of the University of Cambridge, we help millions of people learn English and prove their skills to the world.

For us, learning English is more than just exams and grades. It's about having the confidence to communicate and access a lifetime of enriching experiences and opportunities.

With the right support, learning a language is an exhilarating journey. We're with you every step of the way.

Cambridge Assessment English
The Triangle Building
Shaftesbury Road
Cambridge
CB2 8EA
United Kingdom

 cambridgeenglish.org

 [/cambridgeenglish](https://www.facebook.com/cambridgeenglish)

 [/cambridgeenglishtv](https://www.youtube.com/cambridgeenglishtv)

 [/cambridgeeng](https://twitter.com/cambridgeeng)

 [/cambridgeenglish](https://www.instagram.com/cambridgeenglish)



2015

All details are correct at the time of going to print in November 2019.

Copyright©UCLES 2019