BRITISH COUNCIL

ENGLISH LANGUAGE
ASSESSMENT RESEARCH GROUP

Technical Report
Linking the Aptis Reporting Scales to the CEFR

TR/2015/003

Barry O'Sullivan
British Council

CONTENTS

TABLES

FIGURES

EXECUTIVE SUMMARY

Background

The Aptis development project marked a new era for the British Council, even though it had been involved in a number of test development projects in the past, most notably ELTS (later IELTS). At an early stage in the project, the decision was taken that the test should reflect best practice in the area of language testing and also 'fit' with the British Council's ambitions in the area of assessment literacy. These ambitions relate to the aim of offering world-class advice and consultancy to the many governments, institutions and corporation it works with across the globe. To make the most of the opportunities offered to the British Council itself and to its many partners in the UK and beyond, a wide-ranging assessment literacy agenda has been envisaged in which all British Council staff will be given the opportunity to learn about assessment. In addition, the plan is to pass on this knowledge and expertise to clients so that they can begin to make more informed decisions when it comes to assessment.

Aptis was developed as a low to medium stakes test to be used by large institutions such as education ministries, recruitment agencies and corporations in a variety of situations, where an accurate, though affordable, estimation of the language levels of their employees or prospective employees was required. The decision to undertake a formal CEFR linking project, normally the domain of high stakes tests, reflected a will to continue to push the boundaries of language testing.

The success of the project, as presented in this report, should not be taken as an end in itself. As already indicated, the British Council is committed to a long-term exploration of issues around the validation of Aptis and any future tests it is involved with.

Approach taken

The approach used in the project was based on the procedures recommended by the Council of Europe in their Manual (2003, 2009) and updated by O'Sullivan in the City & Guilds 'Communicator' linking project (2009). As with the design of the whole Aptis development project, the underlying theoretical model of validation is the O'Sullivan & Weir (2011) and O'Sullivan (2011) updated version of the earlier Weir (2005) model.

The approach taken by O'Sullivan (2009) included five stages. However, as the Aptis test was newly developed and much work had already gone into trialling and analysing the various papers, it was decided to eliminate the first stage (critical review) and instead include a review and discussion of the various papers during the standard-setting stage. For this reason, the four stages were:

1. Familiarisation
2. Specification
3. Standard setting
4. Validation

Summary of the main findings

The project findings can be summarised as follows:

1. The Aptis papers offer a broad measure of ability across the different skills, as well as the key area of knowledge of the system of the language.
2. The Aptis test papers are robust in terms of quality of content and accuracy and consistency of decisions.
3. The CEFR boundary points suggested are robust and accurate.

Professor Barry O'Sullivan, British Council
January 2013

1.0 | BACKGROUND TO THE STUDY

The Common European Framework of Reference for Languages (CEFR) was launched over a decade ago by the Council of Europe (2001). Since then it has been translated into approximately 30 languages. It has become the most commonly referenced document upon which language teaching and assessment has come to be based, both in the European Union member states and internationally. An example of its international use is in Taiwan (Wu & Wu, 2010, p. 205), where all nationally recognised examinations must demonstrate a link to the CEFR, and indeed Wu & Wu (2010) describe how the reading paper from a major Taiwanese test (the General English Proficiency Test) was linked to the CEFR. The CEFR provides a description of what it is to know a language from inception to mastery. The descriptors of ability which together comprise the CEFR indicate the progression to mastery and are presented on a rising six-level scale (A1, A2, B1, B2, C1, C2).

The study reported here forms part of a major test development project undertaken by the British Council between 2010 and 2012. The project, which is described in more detail in Section 1.2 below, entailed the development of a test of English language general proficiency covering the CEFR levels A1 to B2 (though also reporting a pre-A1, or A0 level, and a broad C level indication). The test, Aptis, consists of five papers:

1. Grammar & vocabulary	– Language knowledge
2. Reading	– Language use
3. Listening	
4. Writing	
5. Speaking	

1.1. The purpose of the project

As can be seen from the list above, the first paper (known as the 'core' as it always forms part of any Aptis administration) offers a measure of a candidate's knowledge of the systems of the language, focusing specifically on grammar and vocabulary. While performance on the paper is reported on a scale of 0 – 50 (as is the case with the other papers), no CEFR level is applied. This is because the CEFR does not attempt to define levels at the micro level so it would be somewhat meaningless to attempt to determine a link.

The remaining four papers (which focus on language use across the four skills) are reported on the 0 – 50 scale and as a CEFR level. This study was designed to establish empirical evidence of the link between the cut-points on the scale and the claimed CEFR level.

It is important to note that this report describes recommendations made by the standard-setting panel during the final construction stage of the developmental process, at the same time as field trials. These recommendations will need to be revisited in the light of field-trial and operational data analysis and ongoing research and validation. Where appropriate, the cut-scores will need to be revised to take account of what is learnt from these research activities.

1.2. Aptis

As indicated above, Aptis is made up of five language papers, one focusing on a candidate's knowledge of the systems of the language and the others focusing on the ability to actually use the language.

The individual papers are outlined in the following tables (Table 1.1. to 1.5.). A fuller description of the papers, with exemplar tasks, can be found at the Aptis website (www.britishcouncil.org/exams/aptis).

Table 1.1: Overview of the grammar and vocabulary paper

Focus	Task	Format
Grammar	Complete a sentence or utterance.	Three-option multiple choice.
Vocabulary:	Multiple: A. Word definition B. Synonym C. Collocation	Match word to a definition, synonym or collocant. Sets of five target words with ten options.

Table 1.2: Overview of the reading paper

Skill & focus	Task	Format
Reading 1: sentence level comprehension	Series of sentences presenting selective deletion cloze – each sentence is free-standing but appears to form a text.	Five, three-option multiple choice questions, focusing on grammar and vocabulary.
Reading 2: short text cohesion	Re-order a series of sentences to form a story.	Ordering task
Reading 3: short text comprehension	Short text forming a cloze, words to be selected from list.	Text completion using appropriate lexis.
Reading 4: long text overall comprehension	A seven paragraph text given, with series of heading to be matched to each paragraph (with distractors).	Matching

Table 1.3: Overview of the listening paper

Skill & focus	Task	Format
Listening 1: phoneme and word level recognition	Listen short input (such as a phone message) to identify specific information at the phoneme or word level.	Four-option multiple choice for each item (may be listened to twice).
Listening 2: literal meaning	Listen twice to short conversations with two speakers or to monologues to identify specific information.	Four-option multiple choice for each item (may be listened to twice).
Listening 3: inference meaning	Listen twice to short conversations with two speakers or to monologues to identify speaker attitude, intention, mood etc.	Four-option multiple choice for each item (may be listened to twice).

Table 1.4: Overview of the writing paper

Skill & focus	Task	Format
Writing 1: basic word level writing	Complete basic personal information on a type of form.	Form completion
Writing 2: writing two short informal texts	Form completion – two additional personal information questions.	20–30 words each
Writing 3: write three short responses to written input	Respond to input on social network-type website.	Approx. 40 words each
Writing 4: formal and informal text writing	Write one informal message to a friend and a more formal complaint both on the same topic.	Approx. 50 words for Part 1 Approx. 120–150 words for Part 2

Table 1.5: Overview of the speaking paper

Skill & focus	Task	Time
Speaking 1: personal information giving	Respond to three questions, all on everyday, concrete topics.	30 seconds per question
Speaking 2: basic description of picture and comparison with own situation	Picture of concrete event – increasing in complexity (from description to speculation).	45 questions for each question
Speaking 3: describe, compare and speculate	Two contrasting pictures presented, three increasingly complex questions on the two pictures.	Q1 – 40 seconds Q2 – 60 seconds Q3 – 60 seconds
Speaking 4: discuss personal (abstract) ambition, achievement etc.	Picture prompt – though picture is not central to answering the task. Three questions increasing in complexity.	Q1 – 40 seconds Q2 – 40 seconds Q3 – 40 seconds

Not all papers need to be taken by all test takers. Aptis offers a range of packages, from which a client can choose to suit their needs. Table 1.6. shows the range of packages available at launch in August 2012.

The approach taken to the development of Aptis is outlined in O'Sullivan (2015).

Table 1.6: Aptis options

Number	Core (G&V)	Reading	Listening	Writing	Speaking
1	★	★			
2	★		★		
3	★			★	
4	★				★
5	★	★	★		
6	★	★		★	
7	★	★			★
8	★		★	★	
9	★		★		★
10	★			★	★
11	★	★	★	★	
12	★	★	★		★
13	★		★	★	★
14	★	★		★	★
15	★	★	★	★	★

1.2.1. Intended test population

Aptis is designed to assess the English proficiency of non-native speakers of English at CEFR levels A1 to C. It has been designed to be used across a range of contexts and a number of domains, where a measure of general proficiency is required. Aptis is what is called a B-to-B (business-to-business) test, designed to be sold to an institution rather than an individual. It does not offer an internationally recognised certification of ability, but can be certified by the client institution.

Therefore, Aptis is designed to be used where an institution wishes to establish an estimate of the language ability of a known population (e.g. employees, students). The test is designed primarily for adults and young adults. While Aptis was not specifically developed for use with younger learners, it has been shown in controlled trials to function well down to 13 years of age in specific contexts. The British Council expects that potential users with this range in mind carry out carefully designed feasibility studies in cooperation with the British Council to establish empirically that its use is justified.

1.2.2. Stakes and decisions

Aptis is a medium stakes test, designed to allow institutions to make decisions about test takers within that institution. It is not intended for use for high stakes decisions such as university entrance, immigration or citizenship.

2.0 | OVERVIEW OF THE STUDY

The study reported here is comprised of a series of four standard-setting events, each one aimed at establishing a series of cut-points on a separate skill paper. Since the cut-points are designed to indicate different CEFR levels, then a formal standard-setting event is required in order to supply empirical support of the veracity of claims made by the British Council in this regard.

As this is not meant to be a full and formal 'linking' study (as was the case with the City & Guilds Communicator project, O'Sullivan, 2009), it was not considered necessary to follow the complete set of procedures as laid out in the Council of Europe Manual (2009). This is because Aptis is a new testing service, which has been developed from scratch by a team of developers at the British Council using the British Council/ EAQUALS Core Inventory as its basis. The Inventory itself represented a significant attempt to add detail to the CEFR level descriptors by the two organisations. The approach adopted by O'Sullivan (2009) added a critical review phase to the Manual procedure and also limited the claims made with regard to the test at each stage of the process. O'Sullivan also argued that the entire process was iterative, and not linear as inferred by the Manual, and that the process should be supported by a clearly stated model of test validation.

Therefore, the original approach (see Figure 1.1) was replaced with a slightly updated and contextualised version (see Figure 1.2).

Figure 1.1: Model for linking a test to the CEFR (based on O'Sullivan, 2009)

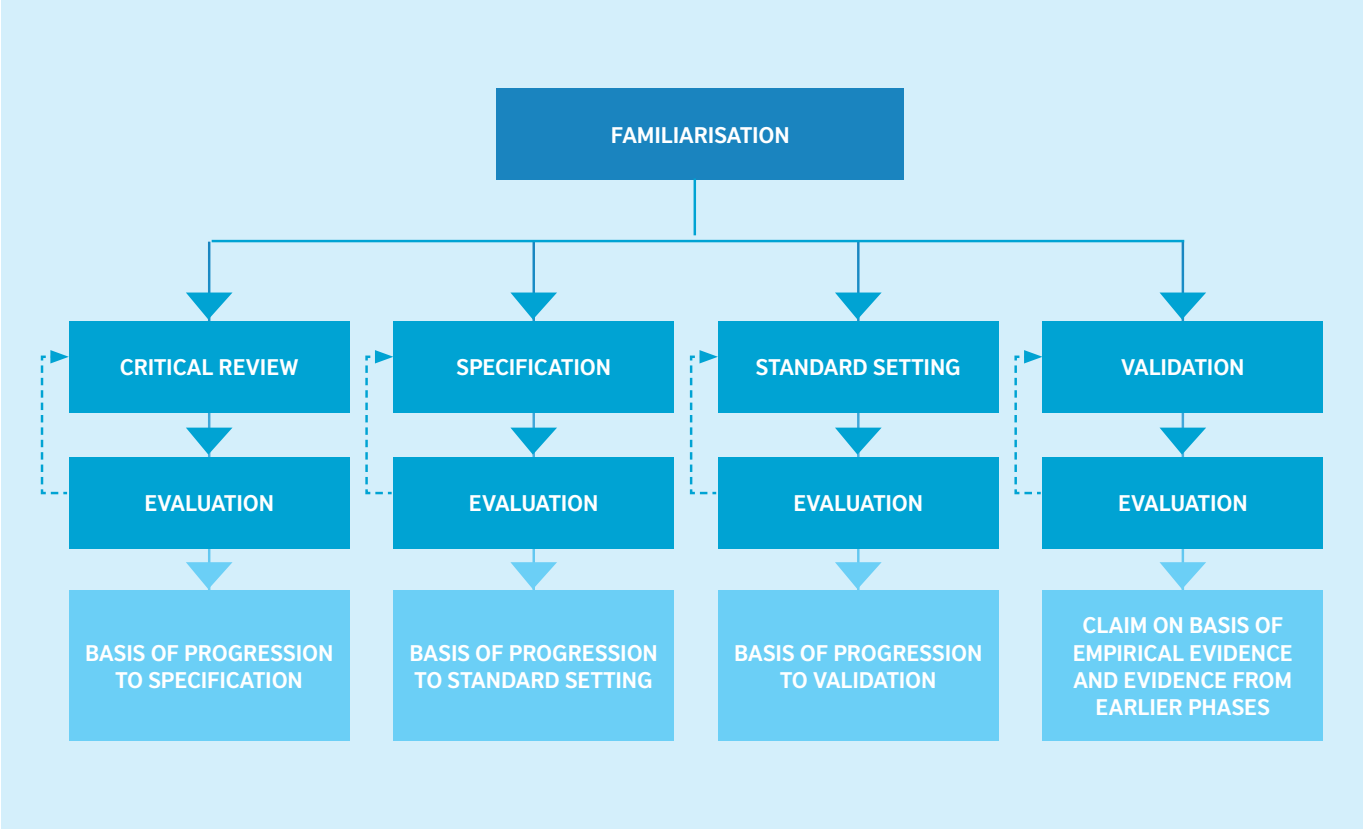
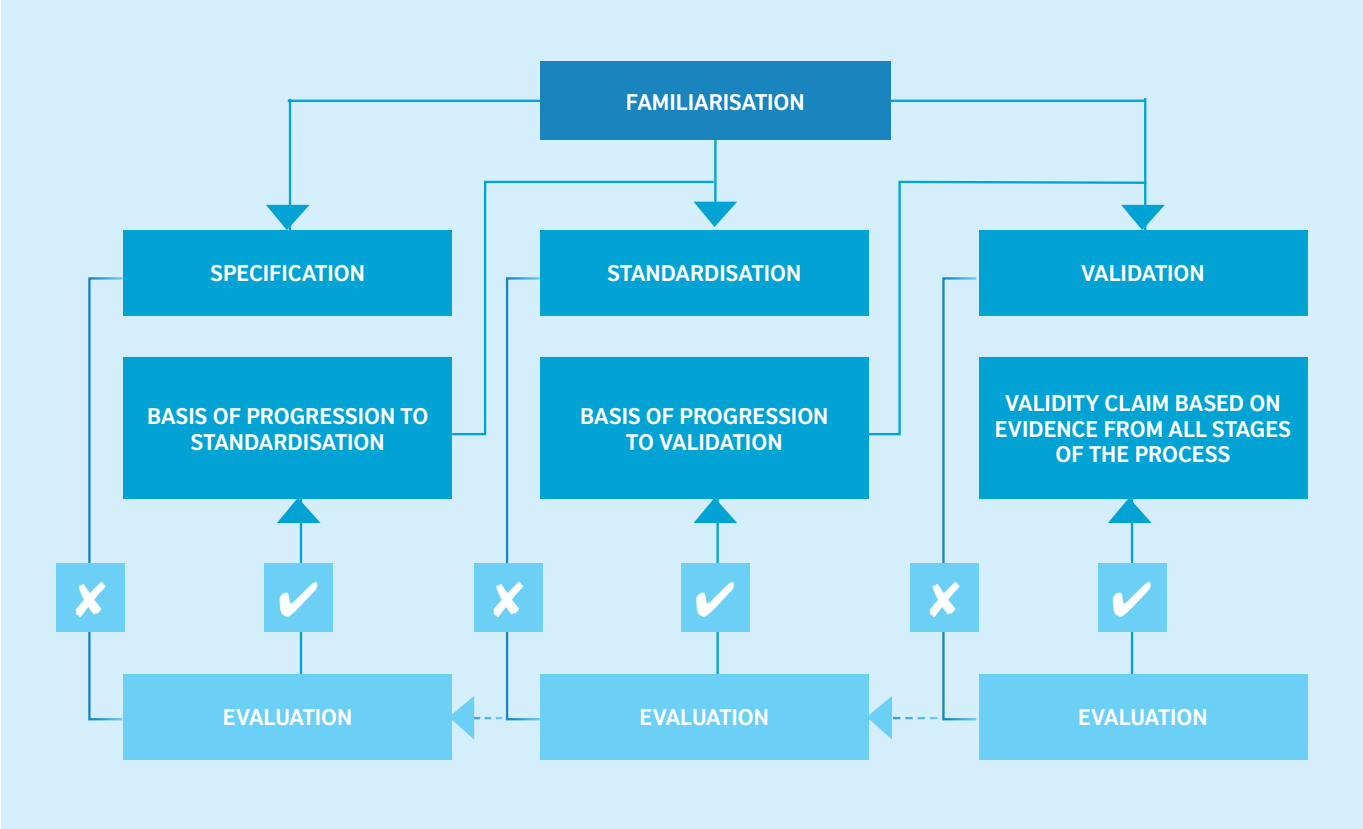


Figure 1.2: Model for linking Aptis to the CEFR



In this model, familiarisation of participants with the CEFR is suggested before each stage of the process. This is to ensure that the participants at these stages are fully competent in their understanding of the purpose of the stage, as well as being able to accurately apply their shared understanding of the CEFR levels.

Once a process has been carried out, it is evaluated in a number of appropriate ways, so that one of two decisions can be made: (1) continue to the next phase (the positive ✔ direction) of the project or (2) go back to the beginning of the stage (or even further back, depending on the findings of the evaluation) and repeat the process, having taking into consideration negative aspects of the evaluation (the negative ✖ direction). Since there are three distinct stages, and as progress is always dependent on the positive outcomes of the evaluation, the entire process can be seen to be iterative in nature.

3.0 | THE SPECIFICATION PHASE

The approach taken by the British Council to creating a set of detailed specifications for the Aptis test papers is unique. This is because, when the system was being devised, it was decided that the danger with creating individual paper-based specifications for different audiences (Alderson, Clapham & Wall, 1995) was that there may well be some changes of focus or emphasis over time which will not be reflected in all these documents. The most critical of these issues would certainly emerge where there are differences between the main specification document and the item writer guidelines. This would result in a very serious validity issue, where the developers and the items writers had differing expectations of the test. To avoid this, a single, web-based specification, based on a custom-made wiki (content management system) was developed. This offers item writers access to the most up-to-date version of the specification, as well as providing a two-way conduit for discussion about the test papers. The system has proven to be highly successful, from the training stage of an item writer's work on Aptis to their day-to-day work.

As the test was planned to link directly to the CEFR, familiarisation activities, based mainly on matching descriptors to CER levels, were combined with review of the CEFR levels as part of the early training of all participants.

In addition, the same validation model with which it was planned to base a future validation argument (used here in the validation phase) was used as the basis of the specifications. This approach was devised by O'Sullivan for the QALSPELL project (Qalspell, 2004) and has been applied in a range of other test development contexts, for example the Zayed University preparatory English programme assessment system in the United Arab Emirates (O'Sullivan, 2005), the EXAVER test in Mexico, Abad et al. (2011) and the COPE test in Turkey (Kantarcıoğlu et al, 2010).

To satisfy the requirements of a Council of Europe linking project, the specification tables provided in the Council of Europe Manual (2009) were completed during the development phase. The completed forms are included here as Appendix 1.

As can be seen in the completed forms, the Aptis papers have been developed to the highest international standards and they assess a broad spectrum of a candidate's language. There is certainly sufficient evidence here to merit continuing to the standard-setting phase of the project.

4.0 | THE STANDARD-SETTING PHASE

In this section of the report, the outcomes of the standard-setting stage for each of the four skills are reported. Since the aim of the familiarisation phase is designed to allow participants in the process to internalise relevant details and interpretations of the CEFR descriptors, this was the first element of each standard-setting event. This element was followed by a discussion of the minimally competent candidate and then rounds of judgements related to items or performances where appropriate.

4.1. The approach taken

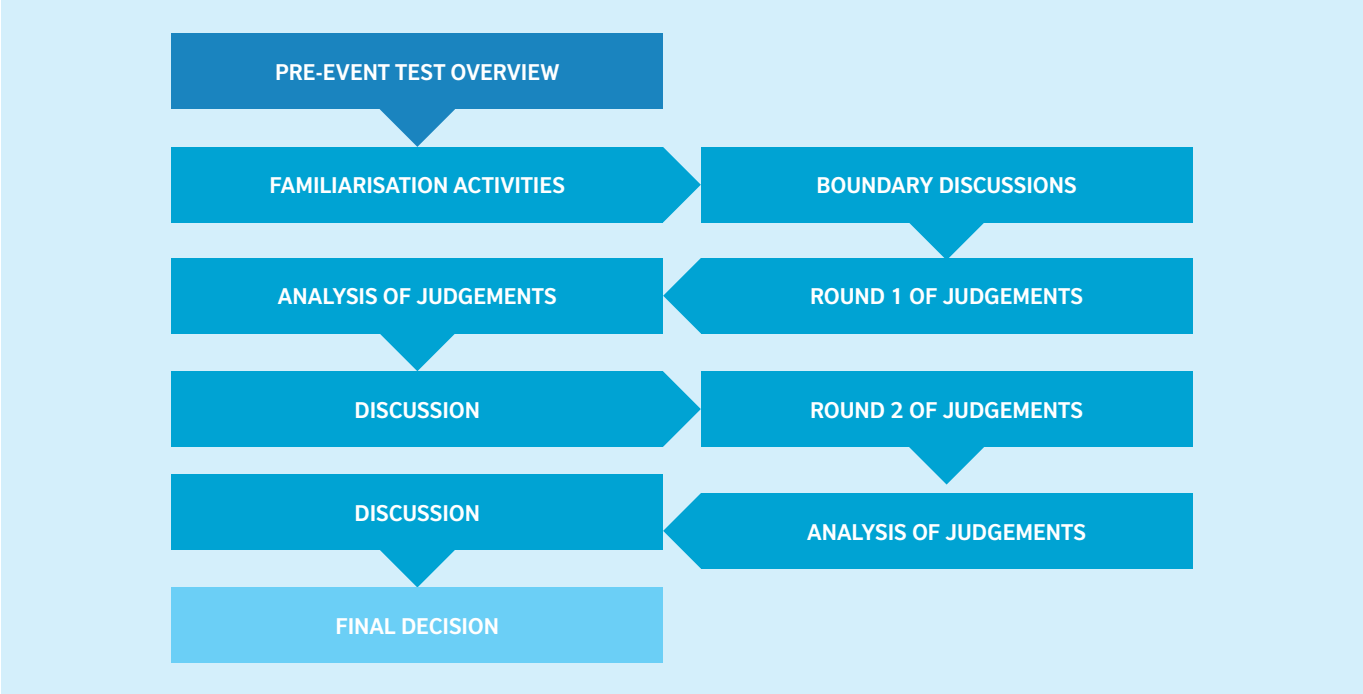
Because of the design of the standard-setting section of the larger Aptis development project, it proved difficult to ask the expert panel participants to undertake a significant amount of pre-event preparation. This was because of the geographic spread of the participants, who were based both inside and outside the United Kingdom. Nevertheless, all participants were sent a pre-event package containing key information about the test.

For the two receptive skills papers (reading and listening), the procedure is shown in Figure 4.1. The participants' aim was to set cut scores for each of the following boundaries:

- A0 – A1
- A1 – A2
- A2 – B1
- B1 – B2
- B2 – C

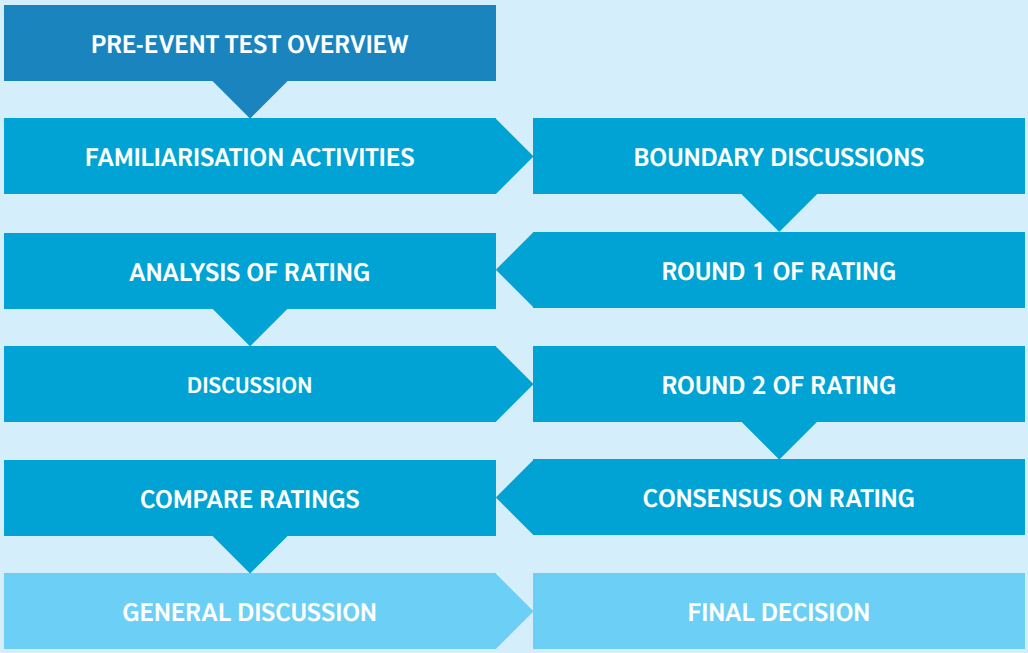
The approach to standard-setting for the receptive skills was a modified Angoff as it offers "the best balance between technical adequacy and practicability" (Berk, 1986, p. 147) and because it has replaced the original Angoff procedure, and is commonly used for tests comprised predominantly of multiple choice (Cisek and Bunch, 2007, p. 82) or other dichotomously scored items. The process is described and summarised in Figure 4.1. and discussed more fully in the relevant sections.

Figure 4.1: Summary of the design of the standardisation process (Knowledge & Receptive papers)



As has been noted elsewhere (O'Sullivan, 2009), when it comes to setting standards for a productive paper (i.e. writing or speaking), the event is more reflective of a rating event than of a typical receptive skill standard-setting event.

Figure 4.2: Summary of the design of the standardisation process (Productive papers)



With the productive papers (speaking and writing), the main focus of our work was to explore the accuracy, in terms of interpretation of level in the Aptis specifications and rater training and standardisation, and of the resultant decisions made by Aptis raters. The procedure for the productive skills is summarised in Figure 4.2 above and outlined in more detail below.

Pre-event test overview — Participants were presented with information about the particular productive skill as described in the CEFR and about the Aptis paper being focused on. The pre-event activity was to review and familiarise themselves with the test paper and re-familiarise themselves with the CEFR level descriptors.

Familiarisation activities — Since the panel selected for the work on the receptive papers was the same as that for the productive papers, we were again able to employ a limited number of familiarisation activities. Again, three such activities were found to be sufficient (though of course many more had been created in case they were needed). The activities were based on matching descriptors to CEFR level. In addition to these activities, panel members were shown a specially constructed scale, created using descriptors from the CEFR, and asked to discuss it and to ensure that it was likely to be functional. By this we mean, the differences between the levels described were clear and easy to distinguish and apply operationally.

Boundary discussions	Following on from the familiarisation activities, participants were asked to discuss the various boundaries, with the aim of internalising the definition of the minimally competent candidate at each boundary point. As was the case with the receptive skills definitions, the resultant definitions are published here. The discussions led to an operational consensus on the levels, their range and the boundaries between them.
Round 1 of judgements	When it was agreed that all participants were ready to begin the rating process, they were asked to consider a set of eight pre-selected scripts. The scripts were selected to represent a range of performances across the CEFR levels and came from different geographical locations (to eliminate any 'local' effects – e.g. a rater might be familiar with the language use or handwriting associated with a particular education system). The panel members then used the scale (described above) to help them decide on the likely CEFR level of each task performance they encountered.
Analysis of judgements	The judgements were entered into a pre-prepared Excel workbook, and individual and group mean CEFR scale levels were automatically estimated. The resulting outputs were then fed back to the participants.
Discussion	Using the data from the first round of ratings, the participants were encouraged to discuss their decisions, particularly where there were significant differences (though in reality there were few if any such cases – with almost all ratings coming within one level of each other). This discussion was led and focused by the event facilitator.
Round 2 of judgement	When the group felt that the discussion had reached a natural conclusion, participants were asked if they wished to reconsider each rating. As was the case with the receptive skills, some participants chose to make changes to their initial ratings based on the preceding discussions, while others did not make any changes.
Analysis of judgements	The data were again entered into the pre-prepared worksheet in the Excel workbook, and the individual and overall mean CEFR levels automatically estimated.
Consensus on ratings	The results of the analysis were then discussed by the participants, who were informed that further rounds of rating could follow if they deemed it necessary, i.e., if they were unable to come to a consensus on the final agreed CEFR level for each task. This option was not required for any of the papers, as overall agreement was quickly reached.
Compare ratings	At this stage in the process, panel members were shown the original scores awarded by Aptis raters (as reported on the CEFR). The idea here was to promote discussion if and when disagreements were found.
General discussion	As it turned out, there was a significant level of agreement between the original levels (from the Aptis raters) and the judgements made by the expert panel.
Final decision	When the final discussion was completed, the participants were asked to agree that the two rating processes had reached an appropriate level of agreement. This was done and the proceedings closed.

4.2. The expert panel

The expert panel for the standard-setting events was made up of 15 individuals, all with extensive teaching and assessment experience, and all with a broad knowledge of the CEFR. None of these individuals had participated in the development of Aptis, and none were currently employed by the British Council. This arrangement was considered to offer the most objective view of the test and of the cut-scores that emerged from the process. Lessons learnt from other attempts at linking language tests to the CEFR include the need for objectivity, reported by O'Sullivan (2009) and the key issue of knowledge of the CEFR, reported by both O'Sullivan (2009) and Kantarcıoğlu (2012).

The outcomes of the standard-setting processes outlined in the previous section are presented in the following pages. Since the general approach has already been described, this section will briefly outline the findings.

4.3. The reading paper

As outlined above (Table 1.2), the reading papers consists of four tasks with a total of 25 associated items. Expert panel members reviewed a complete reading paper when making their judgements.

The expert panel members were asked to participate in an event which was designed to reflect the stages shown in Figure 4.1 above. The details of what occurred during these phases are outlined below.

4.3.1. Pre-event test overview

Participants were presented with information about the reading skill from the CEFR and about the Aptis reading paper. The activity was designed to review and familiarise themselves with the reading paper and re-familiarise themselves with the CEFR level descriptors appropriate to reading.

4.3.2. Familiarisation activities

Since the panel was selected following a competitive process (where a large group applied and selection was based primarily on experience in teaching assessment at a variety of CEFR levels), the initial section of the standard-setting event was set aside for participants to engage in a series of familiarisation activities. Typically three such activities were found to be sufficient (although many more had been created in case they were needed). The activities were based around matching descriptors to CEFR level and they reflect the sorts of tasks suggested in the Council of Europe Manual (2009).

4.3.3. Boundary discussions

Following on from the familiarisation activities, participants were asked to discuss the various boundaries, with the aim of internalising the definition of the minimally competent candidate at each boundary point. In the City & Guilds Communicator Project Report (O'Sullivan, 2009), the definitions were published. However, experience now tells us that the actual definition is, in itself, of little value to anybody who has not participated in the discussion. The real focus is on developing a common understanding of the boundary points and the internalisation of this understanding.

The discussions which led to the operationalization of the boundary points were very positively judged by the panel members, who felt that this section of the event helped them to internalise the definition of the minimally competent reader at each of the boundary points. This, in turn, helped them to identify the criterial differences between the levels and helped make the judging task a lot more efficient. This finding reflects the feelings of the panel members who participated in the City & Guilds project referred to above (O'Sullivan, 2009).

4.3.4. Round 1 of judgements

When it was agreed within the group that all participants were ready to begin the judgement process,
they were asked to consider the whole paper, looking at each item in turn and making two decisions at each boundary point.

1. Is the typical minimally competent candidate (MCC) likely to answer this item correctly (1 = yes; 0 = no)?
2. If there are 100 such people in a test hall, how many are likely to answer the item correctly (steps of 10 between 0 and 100)?

The panel members began the task of judging the reading paper items only when they felt ready to do so following the familiarisation phase. Participants were asked to work alone with no discussion of items with their fellow panel members.

4.3.5. Analysis of judgements from Round 1

Table 4.1. shows a summary of the spreadsheet created for the event – this section focuses only on the
A0 – A1 boundary. As the panel members made their judgements, these were input into the spreadsheet and the suggested cut-score was automatically calculated. In this case, the boundary was found to be at
the 15 per cent point. This automatically calculated boundary was then (together with the other boundary points) presented to the panel as the basis for their discussion of the first round of judgements

Table 4.1: Summary of the Round 1 judgements for the A0 – A1 boundary (Reading)

Items	M	D	A	P	Mi	J	L	An	Ma	Mh	Do	Po	V	N	Ni
1	30	60	40	70	60	50	40	50	40	30	60	10	50	40	50
2	30	30	50	80	60	60	40	60	60	40	60	10	60	60	50
3	20	60	40	70	50	70	40	50	50	50	60	10	50	40	50
4	10	40	40	70	30	80	20	60	50	50	60	30	50	50	50
5	20	20	30	70	20	50	20	60	40	50	50	40	60	50	50
6	10	10	20	40	0	50	40	40	10	20	20	10	50	10	10
7	10	10	10	50	0	50	20	30	10	30	20	10	40	10	10
8	10	10	10	50	0	50	40	40	20	40	20	10	50	10	0
9	10	10	10	50	0	50	20	40	20	40	20	10	40	10	0
10	10	10	10	40	0	50	20	40	30	30	20	10	30	10	0
11	10	10	10	40	0	50	20	30	20	20	20	10	60	10	0
12	10	0	0	0	0	0	10	0	0	0	0	0	0	0	0
13	10	0	0	0	0	0	10	0	0	0	0	0	0	0	0
14	10	0	10	0	0	0	10	0	0	10	0	0	10	0	0
15	10	0	0	0	0	0	10	0	0	0	0	0	0	0	0
16	10	0	0	0	0	0	0	0	0	10	0	0	10	0	0
17	10	0	0	0	0	0	10	0	0	0	0	0	0	0	0
18	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

4.3.6. Discussion of Round 1

The discussions began with the panel members talking about the test tasks, before progressing to the individual items (we consider a task to represent a set of items which are based on a reading input – this can be one or more texts). Discussion of each item was typically started by asking individuals with very different views of the difficulty of that item to try to lead. So, for example, the participants "P" and Po" in Table 4.1 would have been asked to lead the discussions on item 1. The structure of the test meant that the later items were likely to be well beyond the ability of the A1 candidate – a situation broadly recognised by the panel members. When all items had been adequately discussed, we progressed to the next stage of the event.

4.3.7. Round 2 of judgements

When the group felt that the discussion had reached a natural conclusion, participants were asked again to consider each item, asking the same questions as they had done in the first round of judgements. At this point, some participants chose to make changes to their initial judgements based on the preceding discussions, while others did not make any changes. Panel members were asked to reflect on their Round 1 judgements but it was emphasised that they were not to make any changes to their decisions that had not been influenced by the preceding discussion. This was done to avoid any group effect on the judgement process, as it is clear that we are not expecting total unanimity in the judgments of panel members who bring a variety of expertise to the event.

Table 4.2: Summary of the Round 2 judgements for the A0 – A1 boundary (Reading)

Items	M	D	A	P	Mi	J	L	An	Ma	Mh	Do	Po	V	N	Ni
1	40	60	20	30	50	30	40	50	40	30	30	10	30	40	30
2	40	30	30	50	50	60	40	60	60	50	40	10	50	60	50
3	30	60	20	30	50	70	40	50	50	50	40	10	30	40	50
4	20	40	20	40	40	80	20	50	50	50	40	30	40	50	50
5	20	20	20	30	40	50	20	50	40	50	30	40	30	40	50
6	20	10	10	20	20	20	40	40	10	20	20	10	20	10	20
7	20	10	10	30	20	30	20	30	10	30	20	10	30	10	20
8	20	10	10	30	20	30	40	40	20	40	20	10	30	10	0
9	20	10	10	30	20	30	20	40	20	40	20	10	30	10	0
10	20	10	10	20	20	20	20	40	30	30	20	10	30	10	0
11	20	10	10	20	20	20	20	30	20	20	20	10	30	10	0
12	20	0	0	0	0	0	10	0	0	0	0	0	0	0	0
13	20	0	0	0	0	0	10	0	0	0	0	0	0	0	0
14	20	0	10	0	0	0	10	0	0	10	0	0	0	0	0
15	20	0	0	0	0	0	10	0	0	0	0	0	0	0	0
16	20	0	0	0	0	0	0	0	0	10	0	0	0	0	0
17	20	0	0	0	0	0	10	0	0	0	0	0	0	0	0
18	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

4.3.8. Analysis of judgements from Round 2

The data were input into the appropriate section of the spreadsheet and the individual and group cut-scores were again automatically calculated. These data were then used as the basis of the discussion that followed. The summary results for the A0 – A1 boundary point for this round of judgements are shown in Table 4.2. In total, 20 per cent of the judgements were changed by the panel members. The boundary point shifted slightly (to 14 per cent) from Round 1.

4.3.9. Discussion of Round 2

The results of the analysis of the judgements from Round 2 were then discussed by the participants. Again, the initial focus was on items where there tended to be disagreement, though later the individual mean judgements were also discussed to ensure that the group was aware of the impact of the judgements they had made. During this discussion, the panel members were assured that additional rounds of judgements could follow if they deemed it necessary. These additional rounds were not required as the group felt that a final set of decisions could be agreed on.

4.3.10. Final decision

The event concluded with the reaching of a consensus on the placing of the various CEFR boundaries. In fact, the final agreed boundaries reflected those initially identified in Round 1 (the only material change between Rounds 1 and 2 came at the A0 – A1 boundary and that was just a single percentage point lower).

4.3.11. Commentary

The standard-setting event for Aptis Reading resulted in a set of robust boundary points that were endorsed by the expert panel. The experience in reaching a consensus on these boundary points was to contribute to the later panels, as it demonstrated the need for a clearly focused set of test and CEFR familiarisation tasks. It also demonstrated to all concerned (both the organisers and the panel members), the importance of a highly skilled and experienced panel whose members were independent of the test developer.

Following this event, we were satisfied that the boundary points identified represent a genuine and successful attempt to allow Aptis to present test performance data either using a traditional reporting scale and the CEFR (or indeed using both approaches).

4.4. The listening paper

As with the reading paper, the listening paper consists of 25 items. Unlike the reading paper, the listening paper consists of all independent items (while the reading has 25 items focused on four individual reading tasks). The panel members were asked to review a complete listening paper when making their judgements.

The stages shown in Figure 4.1 above, and which were followed in the reading panel, were also followed by the listening panel.

4.4.1. Pre-event test overview

Participants were given details of the listening skill from the CEFR and also of the Aptis listening paper. A set of self-access familiarisation activities were developed to give the panel members an opportunity to familiarise themselves with the listening paper and re-familiarise themselves with the CEFR level descriptors appropriate to listening.

4.4.2. Familiarisation activities

Since the panel for the listening paper consisted of the same members as that for reading paper and the panel met after the reading paper event, we were satisfied that the members were quite familiar with the CEFR level descriptors. However, we insisted on including an opportunity for members to ensure that they really were fully au fait with both the CEFR level descriptors and with the Aptis listening paper by starting the event off with a series of familiarisation activities. As was the case with the reading paper, three familiarisation activities, based on matching descriptors to appropriate CEFR levels, were found to be sufficient.

4.4.3. Boundary discussions

The discussion of the boundaries between the CEFR levels was designed to ensure that the panel shared a common understanding of the differences between the various adjacent CEFR levels. As was the case with the reading event, the definition of these boundaries was meant to offer an operational definition of the broad spectrum of ability and is, therefore, not published here as it is meaningful only to those who participated in the event.

4.4.4. Round 1 of judgements

When the panel members felt that they were ready, the decision was made to progress to the item judgement phase of the event. As with the reading paper, the judgement process was to make the same two decisions at each of the boundary points.

1. Is the typical the MCC likely to answer this item correctly (1 = yes; 0 = no)?
2. If there are 100 such people in a test hall, how many are likely to answer the item correctly (steps of 10 between 0 and 100)?

Again, reflecting the procedure from the reading paper, the participants first worked alone with no discussion of items with their fellow panel members.

4.4.5. Analysis of judgements from Round 1

A spreadsheet was put together for the listening event and used in a similar way to the reading paper. Data from the judgements were entered and the boundary points identified. Table 4.3 shows a summary of the spreadsheet created for the listening event – the section shown here focuses only on the A2 – B1 boundary. After the first round of judgements, the boundary in question was calculated to be at the 43 per cent point. Again, reflecting the approach taken in the reading paper session, the data from the first round of judgements served as the basis for the ensuing discussion.

4.4.6. Discussion of Round 1

The discussions included exploration of both the individual tasks and their difficulty. We approached the discussion in a similar fashion to that of the reading paper, focusing first on individuals with very different views of the difficulty of that item. As with the reading paper, most discussions focused around items where judges were in most disagreement, although all items were touched upon. When the discussion reached a natural break, we progressed to the next stage of the event.

Table 4.3: Summary of the Round 1 judgements for the A2 – B1 boundary (Listening)

Items	M	D	A	P	Mi	J	L	An	Ma	Mh	Do	Po	V	N	Ni
1	80	70	70	100	100	100	90	100	100	80	100	100	70	100	100
2	60	60	100	100	100	100	80	100	100	80	100	100	80	100	100
3	50	70	80	90	100	100	80	100	100	70	100	100	60	100	100
4	70	100	80	70	70	80	90	100	100	80	90	90	40	90	80
5	50	40	50	70	70	50	60	90	100	50	50	90	30	50	80
6	60	60	80	80	100	50	80	100	100	50	30	100	50	100	70
7	30	40	70	60	100	80	60	80	90	40	70	90	30	80	70
8	50	30	50	30	100	20	40	80	80	50	80	50	20	50	50
9	40	40	50	20	90	60	60	60	60	40	80	20	10	50	30
10	30	60	70	50	90	80	60	60	80	40	50	50	30	60	60
11	20	50	40	60	90	80	40	50	70	50	30	30	10	40	50
12	20	30	50	10	80	60	80	50	60	60	20	40	0	30	50
13	30	60	40	0	70	30	60	50	70	70	20	10	20	40	70
14	30	70	60	20	100	60	60	50	60	50	30	60	50	50	50
15	20	50	30	20	60	30	40	40	40	10	10	20	50	50	60
16	30	90	60	10	60	20	60	50	70	60	70	40	0	60	40
17	20	30	40	0	50	20	60	30	50	40	10	0	0	40	50
18	10	50	30	30	50	20	60	30	40	30	60	0	0	40	40
19	50	40	50	20	40	70	60	10	50	30	80	10	50	40	60
20	20	30	30	20	40	30	60	20	30	10	50	0	0	30	50
21	20	10	40	0	50	20	10	20	30	10	40	0	50	20	60
22	10	0	20	0	40	0	10	0	10	0	10	0	0	30	30
23	20	30	40	20	30	10	20	10	20	10	40	10	50	50	50
24	30	40	30	0	20	10	20	10	10	0	30	0	40	30	30
25	20	10	10	0	0	0	30	0	0	0	10	0	0	10	20

4.4.7. Round 2 of judgements

In this phase, participants were asked again to consider each item, asking the same questions as they had done in the first round of judgements. As can be seen in Table 4.4, some panel members made quite extensive changes to their judgements, while others make few if any. As with round one, it was emphasised that panel members were not to make any changes to their decisions that had not been influenced by the preceding discussion.

Table 4.3: Summary of the Round 1 judgements for the A2 – B1 boundary (Listening)

Items	M	D	A	P	Mi	J	L	An	Ma	Mh	Do	Po	V	N	Ni
1	100	70	70	100	100	100	90	100	100	100	100	100	80	100	100
2	70	60	100	100	100	100	80	100	100	100	100	100	90	100	100
3	60	70	80	90	100	100	80	100	100	100	100	100	60	100	100
4	80	100	80	70	70	80	90	100	100	100	90	90	40	90	70
5	60	40	50	70	70	50	60	90	100	50	50	90	60	50	70
6	70	60	50	80	100	50	80	100	100	50	30	100	50	100	60
7	40	40	50	60	100	80	60	80	90	40	70	90	30	80	60
8	60	30	40	30	100	20	40	80	80	80	80	50	20	50	40
9	50	40	50	20	90	60	60	60	60	40	80	30	10	50	30
10	40	60	60	50	90	80	60	60	80	40	50	50	30	60	50
11	30	50	40	60	90	80	40	50	70	50	30	30	30	40	40
12	30	30	40	10	80	60	80	50	60	70	20	40	40	30	30
13	40	60	40	0	70	30	60	50	70	80	20	30	30	40	40
14	40	70	40	20	100	60	60	50	60	50	30	60	50	50	40
15	30	50	30	20	60	30	40	40	40	10	10	30	50	50	40
16	50	90	50	10	60	20	60	50	60	60	70	40	40	60	20
17	30	30	40	0	50	20	60	30	40	40	10	0	40	40	30
18	30	50	30	30	50	20	60	30	30	30	60	0	40	40	20
19	60	40	40	20	40	70	60	10	40	30	80	10	60	40	40
20	40	30	30	20	40	30	60	20	30	10	50	0	0	30	30
21	40	10	40	0	50	20	10	20	30	10	40	0	50	20	40
22	30	0	20	0	40	0	10	0	10	0	10	0	30	30	30
23	40	30	40	20	30	10	20	10	20	10	40	10	50	50	30
24	40	40	30	0	20	10	20	10	10	0	30	0	40	30	10
25	40	10	10	0	0	0	30	0	0	0	10	0	0	10	10

4.4.8. Analysis of judgements from Round 2

The summary results for the A2 – B1 boundary point for this round of judgements are shown in Table 4.4. Interestingly enough, just as was found in the reading paper for the A0 – A1 border, a total of 20 per cent of the judgements were changed by the panel members. The boundary point shifted slightly (to 44 per cent) from Round 1.

4.4.9. Discussion of Round 2

During the discussion of the outcomes from this round of judging, the panel members (who were aware that additional rounds of judgements could follow if they deemed it necessary) focused on items where the differences were most obvious. After some broad discussion, it was decided that a final set of cut-scores could be agreed.

4.4.10. Final decision

The final agreed boundaries reflected those identified in Round 2 (though the only material changes between Rounds 1 and 2 came at three of the boundaries, in each of which a change of just a single percentage point was agreed).

4.4.11. Commentary

The approach taken in the standard-setting event for Aptis Listening resulted in a set of boundary points that were strongly endorsed by the members of the expert panel. As with the reading paper, we were fully satisfied that the boundary points identified represent an accurate estimate of the link between the Aptis scale and the CEFR, and support our practice of presenting test performance data either using a traditional reporting scale or the CEFR levels (or using both approaches).

4.5. The writing panel event

The procedure for the productive skills was outlined in Section 3.1 above and is presented in more detail in this section. As indicated in Section 3.1, there is a major difference when we are dealing with production-based papers, as we are not expecting expert panel members to make judgements on the probability of the difficulty of an item. Instead, we are asking panel members to make informed judgements on the actual level of a piece of language (be it written or spoken) produced by a learner. While we still go through a similarly systematic set of procedures, the task we asked panel members to perform is more akin to rating than to judging, as will be seen.

4.5.1. Pre-event test overview

In the same way as we approached the events which focused on the receptive skills, in the writing event, the expert panel members were presented with information about writing as it is described in the CEFR and also were given detailed information about the Aptis writing paper (task focus, rationale, outcomes and scoring system).

4.5.2. Familiarisation activities

As with the events on the receptive skills, the writing paper standard-setting event began with a orientation phase, in which a number of familiarisation activities were presented to the expert panel members. Three CEFR-related familiarisation activities (though again, more had been created in case they were needed), which involved matching descriptors to specific CEFR levels, were found to be sufficient. These activities were quickly completed by the participants, who demonstrated their familiarity with the CEFR. The development and construction of the writing paper was presented and resulted in some discussion among the members as they internalised all aspects of the paper. In addition to these activities, panel members were introduced to the rating scale. This scale, actually a series of task-specific scales developed for the tasks currently used in the test, are holistic in design and were developed by the Aptis team in conjunction with item-writers and prospective raters. Following extensive trialling, the final versions of the scales were developed prior to the standard setting event. (The scales are shown in Appendix 2.)

4.5.3. Boundary discussions

Before considering a sample of task performances, participants were asked to discuss the various boundaries, with the aim of reaching a consensus on the definition of the minimally competent candidate at each boundary point and internalising these boundaries. The discussions led to an operational definition of the levels, their range and the boundaries between them.

4.5.4. Round 1 of judgements

When all participants were ready to begin the rating process, they were asked to consider a set of eight pre-selected scripts for each of the three tasks. It should be noted here that task 1 in the writing paper (a form completion task aimed at CEFR level A1) does not require a rating scale as it is marked using a set of predetermined rules (e.g. on spelling and punctuation). The scripts were selected to represent a range of performances across the CEFR levels and came from different geographical locations. The panel members then used the scale (see Appendix 2) to help them decide on the likely CEFR level of each task performance they encountered.

The tasks and scripts included in the event are shown in Appendix 3.

4.5.5. Analysis of judgements

The judgements were entered into a pre-prepared Excel workbook, and individual and group mean CEFR scale levels were automatically estimated. The judgements for task 1 are shown in Table 4.5. The level of agreement is quite high except for a small number of tasks and panel members. These disagreements were used as the starting off point for the discussions.

Table 4.5: Round 1 judgements for task 1 (Writing)

Scripts	M	D	A	P	Mi	J	L	An	Ma	Mh	Do	Po	V	N	Ni	Ave Level
93680513	5	5	5	5	5	5	5	5	5	4	5	5	5	5	5	Above A
93680511	5	4	5	5	4	5	4	4	5	3	4	4	5	4	5	A2.2
93680516	5	0	3	4	0	4	2	2	4	3	3	2	3	0	5	A2.1
93680572	4	3	2	3	3	2	3	2	4	2	3	2	3	2	4	A2.1
93683074	3	2	4	3	3	3	2	2	3	2	3	4	3	2	3	A2.1
93683062	4	1	3	2	4	3	2	2	4	1	2	4	2	1	4	A2.1
93683094	5	3	4	3	4	3	2	3	4	1	3	4	4	3	5	A2.1
93683092	5	4	5	4	4	4	5	4	5	3	3	5	5	4	5	A2.2
CofE1	3	3	5	3	3	3	4	3	4	3	3	4	4	2	3	A2.1

The outcomes from the judgements for task 3 are shown in Table 4.6 and are similar in nature to the situation with the first task. Clearly, there was some significant disagreement among the panel members. This was again later used to fuel the discussions.

Table 4.6: Round 1 judgements for task 3 (Writing)

Scripts	M	D	A	P	Mi	J	L	An	Ma	Mh	Do	Po	V	N	Ni	Ave Level
93680513	5	3	4	5	4	4	5	4	4	4	3	5	4	3	5	B1.2
93680511	5	5	5	5	4	5	3	4	5	4	4	5	5	4	5	Above B1
93680516	5	2	4	3	3	1	0	3	5	1	2	1	3	3	5	B1.1
93680572	4	3	2	2	2	3	1	2	3	1	2	1	2	2	4	A2.2
93683074	4	3	2	2	2	2	0	2	3	1	2	1	2	2	4	A2.2
93683062	4	5	5	4	4	5	2	3	4	3	3	3	5	5	5	B1.2
93683094	4	2	2	2	2	3	1	3	3	2	2	1	3	2	3	A2.2
93683092	2	4	4	2	2	3	3	3	3	3	2	3	2	3	4	B1.1
CofE1	3	3	3	4	3	3	2	2	3	4	3	3	0	4	4	B1.1

Finally, the outcomes for task 4 are shown in Table 4.7. Here again there is some level of disagreement, though it does not appear to be as great as with the other tasks.

Table 4.7: Round 1 judgements for task 4 (Writing)t

Scripts	M	D	A	P	Mi	J	L	An	Ma	Mh	Do	Po	V	N	Ni	Ave Level
93680513	5	4	3	3	3	4	5	3	3	3	2	3	4	3	5	B2.2
93680511	4	3	4	2	2	3	2	3	3	4	2	3	5	2	4	B2.1
93680516	3	3	2	1	2	2	0	3	3	2	2	2	3	1	3	B1.2
93680572	1	1	1	0	1	1	0	2	2	1	1	0	2	0	1	B1.1
93683074	2	0	2	0	0	0	0	2	1	2	1	1	1	0	3	B1.1
93683062	3	3	5	3	0	3	0	3	4	4	4	3	4	4	2	B2.1
93683094	1	0	1	0	0	0	0	1	0	0	1	0	0	0	2	Below B
93683092	2	0	1	0	0	0	1	0	1	1	1	0	0	0	3	B1.1
CofE1	1	3	3	2	2	3	4	2	3	3	3	2	3	2	2	B2.1

4.5.6. Discussion

The data tables from the first round of ratings were used as the basis of the discussions. The point of these discussions was to help the panel members further clarify their thinking, and to help them make decisions about any future ratings. This phase of the process was essentially used to replicate the procedures typical of a rater training event.

4.5.7. Round 2 of judgements

When the panel members were satisfied that the discussions had been successfully concluded, they were asked if they wished to reconsider each rating. As was the case with the receptive skills, some participants chose to make changes to their initial ratings based on the preceding discussions, while others did not make any changes. In total just over six per cent of the initial decisions were changed by panel members in the second round.

4.5.8. Analysis of judgements

The data were again entered into the pre-prepared worksheet in the Excel workbook, and the individual and overall mean CEFR levels automatically estimated. The outcomes for the three tasks are presented here in Table 4.8. to 4.10.

Table 4.8: Round 2 judgements for task 1 (Writing)

Scripts	M	D	A	P	Mi	J	L	An	Ma	Mh	Do	Po	V	N	Ni	Ave Level
93680513	5	5	5	5	5	5	5	5	5	4	5	5	5	5	5	Above A
93680511	5	4	5	5	4	5	4	4	5	3	4	4	5	4	5	A2.2
93680516	4	0	3	4	0	4	2	2	4	3	3	2	3	0	5	A2.1
93680572	4	3	2	3	3	2	3	2	4	2	3	2	3	2	4	A2.1
93683074	3	2	4	3	3	3	2	2	3	2	3	4	3	2	3	A2.1
93683062	4	2	3	2	4	3	2	2	4	3	2	4	2	2	4	A2.1
93683094	4	3	4	3	4	3	2	3	4	3	3	4	3	3	3	A2.1
93683092	5	4	5	4	4	4	5	4	5	3	3	5	5	4	5	A2.2
CofE1	3	3	4	3	3	3	4	3	4	3	3	4	4	2	3	A2.1

Table 4.9: Round 2 judgements for task 3 (Writing)

Scripts	M	D	A	P	Mi	J	L	An	Ma	Mh	Do	Po	V	N	Ni	Ave Level
93680513	5	3	4	5	4	4	5	4	4	4	3	5	4	3	5	B1.2
93680511	5	5	5	5	4	5	3	4	5	4	4	5	5	4	5	Above B1
93680516	5	2	4	3	3	3	2	3	3	2	2	2	3	3	4	B1.1
93680572	3	3	2	2	2	3	1	2	3	1	2	1	2	2	3	A2.1
93683074	4	3	2	2	2	2	0	2	3	1	2	1	2	2	3	A2.2
93683062	4	5	5	4	4	5	2	3	4	3	3	3	5	5	5	B1.2
93683094	3	2	2	2	2	3	1	3	3	2	2	1	3	2	3	A2.1
93683092	2	4	4	2	2	3	3	3	3	3	2	3	2	3	4	B1.1
CofE1	3	3	3	4	3	3	2	2	3	4	3	3	3	4	4	B1.1

Table 4.10: Round 2 judgements for task 4 (Writing)

Scripts	M	D	A	P	Mi	J	L	An	Ma	Mh	Do	Po	V	N	Ni	Ave Level
93680513	5	4	3	3	3	4	5	3	3	3	3	3	4	3	5	B2.2
93680511	3	3	3	2	2	3	2	3	3	4	2	3	3	2	4	B2.1
93680516	3	3	2	1	2	2	1	3	3	2	2	2	3	1	3	B1.2
93680572	1	1	1	0	1	1	0	2	2	1	1	0	2	0	1	B1.1
93683074	2	0	2	0	0	0	0	2	1	2	1	1	1	0	2	B1.1
93683062	3	3	4	3	3	3	2	3	4	4	4	3	4	4	2	B2.1
93683094	1	0	1	0	0	0	0	1	0	0	1	0	0	0	1	Below B
93683092	2	0	1	0	0	0	1	0	1	1	1	0	0	0	2	B1.1
CofE1	2	3	3	2	2	3	4	2	3	3	3	2	3	2	2	B2.1

4.5.9. Consensus on ratings

The results of the analysis presented in the previous section were then discussed by the panel members. As with the reading and listening events, they had been assured that additional rounds of judgements could be added if they felt that this might be necessary. This option was not required for the writing paper, as the group decided after some discussion, that the indicative CEFR levels identified in the second round of judgements were acceptable.

4.5.10. Compare ratings with original

Panel members were now shown the original scores awarded by Aptis raters (as reported on the CEFR) for each of the 24 Aptis tasks. It was immediately clear that there was complete agreement between the levels suggested by the expert panel and those indicated by the Aptis raters.

4.5.11. General discussion

The significance of the level of agreement between the original levels (from the Aptis raters) and the judgements made by the expert panel meant that there was little meaningful general discussion at this stage. The main decision was to end the process as additional judgements were felt to be unnecessary.

4.3.12. Final decision

When the brief final discussion was completed, the participants agreed that the Aptis rating system was working well, in that it was both easy and intuitive to apply and the outcomes were consistent with expectations. The fact that the three Council of Europe recommended tasks (which were claimed to represent levels A2, B1 and B2) were also judged by the panel members to be at the levels claimed (though in all cased towards the lower end of the level) was an additional indication of the accuracy of the Aptis approach.

4.6. The speaking panel event

The procedure for the speaking test was the same as that for writing and is outlined in this section.

4.6.1. Pre-event test overview

The expert panel members were presented with information about speaking as it is described in the CEFR and were also provided with detailed information about the Aptis speaking paper (task focus, rationale, outcomes and scoring system).

4.6.2. Familiarisation activities

Three CEFR-related familiarisation activities, which again were based on a series of matching tasks, linking descriptors to specific CEFR levels, were used. The development and construction of the speaking paper was presented and, in addition, the panel members were introduced to the rating scale. This scale is unlike that of the writing paper, as it is actually a single, test (as opposed to task) specific holistic scale. The scale was developed by the Aptis team in conjunction with item-writers and prospective raters. Following extensive trialling and feedback from users, the final version of the scale was agreed on prior to the standard setting event. The scale is shown in Appendix 4.

The tasks used in the event were from the main Aptis trials and are shown in Appendix 5.

4.6.3. Boundary discussions

As with the standard-setting event for the writing paper, panel members were asked to discuss the various boundaries, again with the aim of internalising these boundaries and on reaching a consensus on the definition of the minimally competent candidate at each boundary point. The resulting operational definition of the levels, their range and the boundaries between them underpinned the later judgements and discussions.

4.6.4. Round 1 of judgements

Participants were then asked to listen to a set of eight pre-selected recordings of candidate performances on the four tasks (see Appendix 5). The audio files represented a range of performances and were rated using the scale shown in Appendix 4.

4.6.5. Analysis of judgements

The judgements for task 1 are shown in Table 4.11. The level of agreement is quite high except for a single panel member (An). Any disagreements were later used as the starting off point for the discussions.

Table 4.11: Round 1 judgements for task 1 (Speaking)

Scripts	M	D	A	P	Mi	J	L	An	Ma	Mh	Do	Po	V	N	Ni	Ave Level
93690131	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Below A
93686854	3	2	2	3	3	2	2	2	3	3	2	3	3	2	2	A2
93686728	3	2	2	3	3	3	3	1	3	2	2	2	2	2	3	A2
93690115	3	4	4	4	2	3	4	2	4	4	3	2	2	3	4	B1
93686711	4	3	3	4	3	4	4	1	4	3	2	2	3	3	4	B1
93686905	3	3	2	4	3	3	3	2	3	3	3	4	2	3	3	B1
93690100	4	4	4	4	4	3	5	3	4	2	3	3	4	4	3	B2
93685911	4	4	3	4	3	3	4	2	4	4	3	3	4	3	5	B2
CofE1	2	3	3	2	2	3	4	2	3	3	3	2	3	2	2	B2.1

Table 4.12: Round 1 judgements for task 2 (Speaking)

Scripts	M	D	A	P	Mi	J	L	An	Ma	Mh	Do	Po	V	N	Ni	Ave Level
93690131	1	1	1	1	2	1	1	1	2	1	1	2	1	1	1	A1
93686854	3	3	3	4	4	3	3	2	4	4	3	3	4	2	4	B1
93686728	3	2	2	3	3	3	3	1	3	3	2	2	2	2	3	A1
93690115	3	4	4	4	2	4	4	2	4	4	3	3	3	3	4	B1
93686711	3	2	3	3	3	4	4	2	3	4	2	2	3	3	4	B1
93686905	2	3	2	4	2	3	3	2	3	2	3	3	2	3	3	B1
93690100	5	4	3	4	4	4	5	3	4	3	3	3	4	4	4	B2
93685911	4	3	3	4	3	4	4	1	3	4	3	3	3	3	5	B1
CofE1	2	3	3	2	2	3	4	2	3	3	3	2	3	2	2	B2.1

Table 4.13: Round 1 judgements for task 3 (Speaking)

Scripts	M	D	A	P	Mi	J	L	An	Ma	Mh	Do	Po	V	N	Ni	Ave Level
93690131	2	1	1	1	2	1	1	0	1	1	1	2	1	1	2	A1
93686854	3	3	3	4	3	3	3	2	4	3	2	3	2	3	4	B1
93686728	3	2	2	3	2	3	3	1	3	3	2	1	3	2	3	A1
93690115	3	3	4	4	3	3	3	2	4	4	3	4	4	3	4	B1
93686711	3	3	3	3	3	4	4	2	3	4	2	2	3	3	4	B1
93686905	2	2	3	4	2	4	3	1	3	2	3	3	2	3	3	B1
93690100	5	4	3	4	3	4	5	2	4	2	3	3	3	4	4	B2
93685911	3	3	3	3	3	2	3	1	3	3	2	3	3	2	4	B1
CofE1	2	3	3	2	2	3	4	2	3	3	3	2	3	2	2	B2.1

Table 4.14: Round 1 judgements for task 4 (Speaking)

Scripts	M	D	A	P	Mi	J	L	An	Ma	Mh	Do	Po	V	N	Ni	Ave Level
93690131	2	1	1	2	2	1	2	1	1	1	1	1	2	1	2	A1
93686854	4	4	3	4	4	3	3	2	4	4	2	3	4	3	4	B1
93686728	3	3	2	3	2	3	3	1	3	3	2	2	3	2	3	B1
93690115	3	4	4	4	3	3	4	2	4	4	3	5	4	3	4	B2
93686711	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	A0
93686905	3	3	2	4	3	3	3	1	4	3	3	4	3	3	3	B1
93690100	5	5	3	5	4	5	5	3	5	3	3	4	4	4	5	B2
93685911	4	4	3	4	3	4	3	2	4	5	2	4	4	3	5	B2
CofE1	2	3	3	2	2	3	4	2	3	3	3	2	3	2	2	B2.1

4.6.6. Discussion

There was a high level of agreement among the panel members, though one ('An') appeared somewhat wayward when compared to the others. That said, the deviation shown by 'An' tended to be in the same direction. In other words, this panel member tended to be consistently harsher than the others when making judgements. The discussions, therefore, centred around these (and other less significant) variations.

4.6.7. Round 2 of judgements

As with the other papers, some participants chose to make changes to their initial ratings based on the preceding discussions, while others did not make any changes. In total, less than three per cent of the initial decisions were changed by panel members in the second round.

4.6.8. Analysis of judgements

The outcomes for the four tasks are presented here in Table 4.15 to 4.18.

Table 4.15: Round 2 judgements for task 1 (Speaking)

Scripts	M	D	A	P	Mi	J	L	An	Ma	Mh	Do	Po	V	N	Ni	Ave Level
93690131	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Below A
93686854	3	2	2	3	3	2	2	2	3	3	2	3	3	2	2	A2
93686728	3	2	2	3	3	3	3	2	3	2	2	2	2	2	3	A2
93690115	3	4	4	4	2	3	4	2	4	4	3	2	2	3	4	B1
93686711	4	3	3	4	3	4	4	3	4	3	2	2	3	3	4	B1
93686905	3	3	2	4	3	3	3	2	3	3	3	4	2	3	3	B1
93690100	4	4	4	4	4	3	5	3	4	2	3	3	4	4	3	B2
93685911	4	4	3	4	3	3	4	3	4	4	3	3	4	3	5	B2
CofE1	2	3	3	2	2	3	4	2	3	3	3	2	3	2	2	B2.1

Table 4.16: Round 2 judgements for task 2 (Speaking)

Scripts	M	D	A	P	Mi	J	L	An	Ma	Mh	Do	Po	V	N	Ni	Ave Level
93690131	1	1	1	1	2	1	1	1	2	1	1	2	1	1	1	A1
93686854	3	3	3	4	4	3	3	2	4	4	3	3	4	2	4	B1
93686728	3	2	2	3	3	3	3	1	3	3	2	2	2	2	3	A2
93690115	3	4	4	4	2	4	4	2	4	4	3	3	3	3	4	B1
93686711	3	2	3	3	3	4	4	2	3	4	2	2	3	3	4	B1
93686905	2	3	2	4	2	3	3	2	3	2	3	3	2	3	3	B1
93690100	5	4	3	4	4	4	5	3	4	3	3	3	4	4	4	B2
93685911	4	3	3	4	3	4	4	3	3	4	3	3	3	3	5	B1
CofE1	2	3	3	2	2	3	4	2	3	3	3	2	3	2	2	B2.1

Table 4.17: Round 2 judgements for task 3 (Speaking)

Scripts	M	D	A	P	Mi	J	L	An	Ma	Mh	Do	Po	V	N	Ni	Ave Level
93690131	2	1	1	1	2	1	1	0	1	1	1	2	1	1	2	A1
93686854	3	3	3	4	3	3	3	2	4	3	2	3	2	3	4	B1
93686728	3	2	2	3	2	3	3	1	3	3	2	1	3	2	3	A2
93690115	3	3	4	4	3	3	3	2	4	4	3	4	4	3	4	B1
93686711	3	3	3	3	3	4	4	2	3	4	2	2	3	3	4	B1
93686905	2	2	3	4	2	4	3	2	3	2	3	3	2	3	3	B1
93690100	5	4	3	4	3	4	5	3	4	3	3	3	3	4	4	B2
93685911	3	3	3	3	3	2	3	2	3	3	2	3	3	2	4	B1
CofE1	2	3	3	2	2	3	4	2	3	3	3	2	3	2	2	B2.1

Table 4.18: Round 2 judgements for task 4 (Speaking)

Scripts	M	D	A	P	Mi	J	L	An	Ma	Mh	Do	Po	V	N	Ni	Ave Level
93690131	2	1	1	2	2	1	2	1	1	1	1	1	2	1	2	A1
93686854	4	4	3	4	4	3	3	2	4	4	2	3	4	3	4	B1
93686728	3	3	2	3	2	3	3	2	3	3	2	2	3	2	3	B1
93690115	3	4	4	4	3	3	4	3	4	4	3	5	4	3	4	B2
93686711	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	A0
93686905	3	3	2	4	3	3	3	3	4	3	3	4	3	3	3	B1
93690100	5	5	3	5	4	5	5	3	5	3	3	4	4	4	5	B2
93685911	4	4	3	4	3	4	3	3	4	5	2	4	4	3	5	B2
CofE1	2	3	3	2	2	3	4	2	3	3	3	2	3	2	2	B2.1

4.6.9. Consensus on ratings

As with the other standard-setting events, panel members had been told that additional rounds of judgements could be added if necessary. These were not required for the speaking paper, as the group decided after some short discussion that the indicative CEFR levels identified in the second round of judgements were acceptable.

4.6.10. Compare ratings with original

Panel members were again at this stage shown the original scores awarded by Aptis raters (as reported on the CEFR) for each of the tasks. It was clear that there was complete agreement between the levels suggested by the expert panel and those indicated by the Aptis raters.

4.6.11. General discussion

The indication of agreement between the Aptis ratings and the judgements made in the standard-setting event confirmed that there would be no need for further judgements or discussion.

4.6.12. Final decision

Finally, as with the writing paper, the participants agreed that the Aptis rating system was easy and intuitive to apply and the outcomes were consistent with expectations.

4.7. Claims

It is now widely accepted that, in order to make any claim of a link between a test's reporting system and the CEFR, it is first necessary to indicate how the test itself has been informed by the CEFR, and then demonstrate how important decision boundaries are made. When deciding on the precise placement of these boundaries, we should again both demonstrate how the decision-making process has been informed by the CEFR and also demonstrate that the process has been transparent, systematic and accurate.

This section of the report has clearly demonstrated that the boundaries set for the Aptis papers are robust and reliable. However, remaining consistent with earlier suggestions related to the linking process (O'Sullivan 2009), we will not at this point be making any final definitive claims regarding Aptis. All final claims will be addressed following the presentation of evidence of the validity of the examination.

The procedures followed in the standard-setting process indicate that the validity of the claim of the CEFR levels reported by Aptis is strong.

5.0 | THE VALIDATION STAGE

To ensure that the final element of the linking project provides a coherent argument, we will present evidence of the validity of Aptis in terms of the components of the validation frameworks developed initially by Weir & O'Sullivan at Roehampton University over a decade ago and later published by Weir (2005) and updated by O'Sullivan & Weir (2011) and O'Sullivan (2011). This model is used here for a number of reasons. As argued by O'Sullivan & Weir (2011), the framework is the only practically operational model of validation in existence. Others have been proposed but they fail to offer the user a sufficiently detailed or coherent account of what is expected of the validation process, see for example Kane (1972), Messick (1975, 1980, 1989) or Mislevy et al. (2002, 2003). The following elements of the updated framework are presented for each paper:

- the test taker
- the test task
- the scoring system

5.1. The test taker

The test taker is considered within the Aptis approach from the very beginning. The underlying model of validation which drives the test is O'Sullivan's (2011) modification of the earlier Weir (2005) validation framework. This model suggests that we consider the test taker from two major perspectives, personal characteristics and cognitive characteristics. The successful test will take both sets of characteristics into account when creating items and tasks.

5.1.1. Personal characteristics

The Aptis approach to the challenge of dealing with test taker personal characteristics is to routinely communicate with clients to ensure that the test is appropriate for the candidature both linguistically and culturally. While the test tasks and items are developed for a general audience, the policy of the British Council is to meet with prospective clients to ensure that the test takers reflect the intended general population. Where this is not the case, for example, when the students are younger than the group the test has been created for, we either suggest an alternative test or propose a research study to investigate test performance from a representative sample of the population. At the time of writing this report, this has been done in three countries. In one of these places (India), the results of a review process resulted in some changes to the test (e.g. people and place names, reading task topics, writing task topic, and speaking task photographs) before it was considered appropriate for trialling with a group of secondary school students (Maghera & Rutherford, 2013). At a trial of the test in these circumstances, various additional pieces of information are collected, these include:

- linguistic background (L1)
- age
- educational level
- ethnic background
- gender

Some of these variables (age and gender) are routinely collected and test data analysed for potential bias.

In addition to these procedures, Aptis is delivered using the Surpass platform (BTL, 2012), which conforms to all international requirements for accessibility.

5.1.2. Cognitive challenge

All Aptis papers are designed to gradually increase the cognitive challenge on the candidate as the test progresses. How this is actually achieved is presented in Table 5.1.

By careful consideration of the cognitive challenge involved in each task and item, Aptis ensures that the test is likely to offer a challenge to candidates of all levels, while still offering the weaker candidates enough to demonstrate their language ability to the full. Together with the parameters discussed in Section 5.2, this approach ensures a sufficiently representative coverage of each skill area.

Table 5.1: Cognitive challenge in Aptis tasks

Listening	Reading	Writing	Speaking
Listening is assessed using 25 discrete items. The items are designed to become progressively more difficult and reflect the different aspects of listening assessed: 1. Sound System 2. Literal Meaning 3. Inferred Meaning (Weir, 1993; Buck 2001)	**Task 1:** Limited challenge, concrete topic and task and simple sentence level understanding.	**Task 1:** Lowest challenge, basic personal information in online form.	**Task 1:** Lowest challenge, short responses to three personal questions, though they increase slightly in complexity.
	Task 2: Higher challenge, identification of cohesive structure, so sentence level focus, though within text.	**Task 2:** Slightly higher challenge, individual preferences (interests etc.) in sentence form.	**Task 2:** Higher challenge as candidate describes picture, then responds to personal question related to picture, then makes comparisons between scene in picture and own culture/city etc.
	Task 3: Slightly higher again, with emphasis on reflecting on whole test.	**Task 3:** Higher challenge, candidate reads input and must react (social media), followed by additional inputs and further reaction (mimics interactive writing).	**Task 3:** Higher again as here two pictures to be described in some way, then compared then speculation (e.g. which is best to visit for a holiday, why?)
	Task 4: Highest challenge, integration of headings into long text, calls for local and global understanding.	**Task 4:** Highest challenge, candidate faced with dilemma and must report casually in email to friend while also reacting formally to person of high status.	**Task 4:** Highest challenge as candidate is presented with abstract topic (illustrated but not supported by picture) and first personalise it, then talk about emotions before finally discussing the topic on an abstract level.

5.2. The test task

We will look at the test task from three perspectives: the setting (or physical parameters); the demands (or linguistic parameters); and the administration. The concept of context validity (Weir, 2005) has been replaced with the notion of the validity evidence which supports the use of a particular test task. This evidence can be seen from the perspectives of task-setting parameters (the conditions under which the task is performed), task-demand parameters (the linguistic demands of the input and expected output), and the conditions surrounding the administration of the test (the non-language aspects of task administration). The meaning of these parameters are summarised in Appendix 6.

5.2.1. The physical parameters

The physical parameters are outlined in Table 5.2 and additional information is presented in the sub-sections that follow.

Table 5.1: Cognitive challenge in Aptis tasks

Parameter	Core	Listening	Reading	Writing	Speaking
Purpose	Assess knowledge of the language system (grammar & vocabulary	General proficiency			
Intended population	Mid-teens to adult learners of English Data collected at test trial and administration stages (completed by each test taker)				
Number of tasks/items	Grammar – 25 items	25 items	4 tasks (25 items)	4 tasks	4 tasks
Response format	Vocabulary – 5 tasks (25 items)	MCQ	Combination of MCQ, re-ordering sentences and matching	Handwritten response	Spoken response
Known criteria	Grammar – MCQ	N/A – answer key	N/A – answer key	Not shown on test, is available online on the Aptis website	
Task types	Grammar – focus on grammatical form (including discourse usage) using MCQ items Vocabulary – Focus on word definition, usage, synonyms, collocations	1. Listening for detail (pragmatic competence) – MCQ items 2. Listening for overall meaning – MCQ items 3. Listening for detail – note taking 4. Listening for detail – MCQ items	1. Careful local reading – MCQ cloze items 2. Careful global reading – re-building text 3. Global meaning –gapped text/ matching 4. Global reading – overall meaning	1. Form filling 2. Short extended guided writing (personal information) 3. Interactive writing (social media semi-guided 4. Extended writing – informal and formal text	1. Personal information questions 2. Short answer non-personal questions (picture prompt) 3. Describe and compare questions (2 picture prompts) 4. Extended output based on prompt
Weighting	Equal weighting	Equal weighting	Equal weighting	Weighted: Task 1: max 3 Task 2: max 5 Task 3: max 7 Task 5: max 9	Equal weighting
Order of Items	Order as above for computer, test takers may respond in any order on P&P	As described in task types	Order as above for computer, test takers may respond in any order on P&P	Order as above for computer, test takers may respond in any order on P&P	As described in task types
Time constraints	30 minutes	25-50 minutes (may listen twice)	40 minutes	40 minutes	15 minutes

Key: MCQ – multiple choice questions; P&P – pen and paper.
Additional notes on key parameters are presented in the following sub-sections.

5.2.1.1. Purpose

In each of the papers in the Aptis system, candidates are offered a wide variety of tasks, each with specifically defined purposes. The rationale behind this approach is to ensure as broad a coverage of the underlying construct as possible, and to ensure that candidates are encouraged to set goals from the beginning of each task that reflect those expected by the development team.

The flexibility of the Aptis approach means the British Council is in a position to work with clients to localise (i.e. make appropriate to the particular context and domain of language uses) the test, thus ensuring it will meet the expectations and requirements of the client while maintaining its internal integrity (from a content and a measurement perspective). An example of this approach was presented by Maghera & Rutherford (2013) when describing the work done on the Aptis papers to ensure they met the needs of a major Indian client.

5.2.1.2. Response format

In the same way that the different items and tasks have a variety of purposes, they also contain a range of response formats, from multiple choice to matching in the knowledge and receptive skills papers, to structured and open responses in the productive skills papers. This commitment to offering a wide variety of task and item formats reduces the potential for any format-related bias (either positive or negative).

5.2.1.3. Known criteria

In order to ensure that all candidates set similar goals with regard to their expected responses, the assessment criteria for all tasks and items are made clear both within the test papers and on the Aptis website.

It is also the case that the assessment criteria were very carefully considered by the development team in the early stages of the process to ensure that they reflect the underlying knowledge and ability being assessed in each paper. This link is recognised by Weir (2005) and O'Sullivan and Weir (2011) as being critical to the validity of the test.

5.2.1.4. Weighting

All items are equally weighted in each paper and this information is made clear to the candidates both within the paper and on the Aptis website. This is done to ensure that candidates are all equally informed as to the expectations of the developers (and therefore do not spend more time than intended on particular aspects of the test).

5.2.1.5. Order of items

While the papers are set out in a particular order, the candidate is free to respond in any order, with the exception of the speaking and the listening papers.

5.2.1.6. Time constraints

Candidates are allowed a limited amount of pre-performance preparation time for both writing and speaking (the time is built into the response times). In addition to this, the time allowed for responding to items and tasks is carefully controlled to ensure a similar test experience for all candidates. In fact, all timings are automatically gathered and will be used by the Aptis research team to study specific aspects of the test papers.

5.2.2. The linguistic parameters

The linguistic parameters refer to the language of the input, the expected output and also to factors such as variables associated with the interlocutor or audience that may affect language performance, e.g. gender, status, nature of acquaintanceship, see O'Sullivan (2000, 2002, 2008) and Berry (2007). These are explained in Appendix 5 and outlined as they relate to Aptis in Table 5.3.

5.2.2.1. Channel

In terms of input, this can be written, visual (photo, artwork, etc.), graphical (charts, tables, etc.) or aural (input from examiner, recorded medium, etc.). Output depends on the ability being tested, although candidates will use different channels depending on the response format. With Aptis, we consider channel from a number of perspectives, taking into account lessons learnt from multi-literacy research (see Unsworth, 2001) and assessment research (Ginther, 2001; Wagner, 2008) into the impact on test performance of features of visual input, for example.

5.2.2.2. Discourse mode

This includes the categories of genre, rhetorical task and patterns of exposition and is reflected in the input (in the receptive skills papers) and the output (in the productive skills papers). A good example of how we approach this is in task 4 of the writing paper. Here, the candidate is asked to respond to a dilemma in a number of ways, for example, in an email to a friend and in a letter of complaint to a person in authority (thus identifying a candidate's ability to recognise register in their written response).

Table 5.3: Test task evidence (demands) of the Aptis papers

Parameter	Core	Listening	Reading	Writing	Speaking
Discourse mode	Grammar: short input, using descriptive, narrative and discursive texts Vocabulary: discrete (single word) and simple descriptive texts	Announcements Phone messages / conversations [formal & informal] Short monologue All including a range of accents All delivered at normal speed (approx. 125-150 words per minute)	Task 1: related sentences – each can be understood in isolation Task 2: narrative or biographical text of 7 sentences. Task 3: medium length narrative or descriptive text of 1 or 2 paragraphs. Task 4: a longer narrative, discursive, explanatory, descriptive, or instructive text.	Task 1: form completion Task 2: personal information exchange. Task 3: informal narrative or descriptive text Task 4: two texts a) informal email b) formal request or complaint	Task 1: response to short personal questions (interactive type) Task 2: descriptive & narrative Task 3: description, comparison & speculation Task 4: to texts – extended monologue
Channel	Written	Aural	Written	Written	Spoken
Text length	Grammar: maximum 15 words for most forms, 30 words for discourse-based items. Vocabulary: typically single word, short sentences (10 word max) for usage items.	Typically approx. 50 word input 1 to 5 word options in MCQs	Task 1: Max 50 words. Task 2: 100 words in 7 sentences. Task 3: 135 words Task 4: 750 words in text. Headings are a maximum 12 words.	Task 1: 110-130 words Task 2: 30-50 words in input text. Maximum 40 words in rubric. Task 3: Maximum 25 words in rubric. 100-120 words written by learner.	Task 1: max. 10 words per question (x3) Task 2: max 25 words Task 3: approx. 100 words Task 4: approx 35 words
Writer-reader relationship	Not relevant to these items as they are accessing an individual's knowledge of the language system and are not concerned with usage, with the exception of discourse appropriacy.	In most cases the speaker is identified as a friend, colleague or boss for example. With other items (e.g. announcements), it is assumed that the speaker is a stranger.	Task 1: friend, family Task 2: unspecified Task 3: unknown writer Task 4: unknown writer	Task 1: unknown reader Task 2: unknown reader Task 3: friend Task 4: a) friend, b) person of status	Unspecified audience for all tasks
Nature of information	Concrete	Concrete and some more abstract	Concrete and some more abstract	Concrete and some more abstract	Concrete and some more abstract

Continues overleaf.

Continued: Table 5.3: Test task evidence (demands) of the Aptis papers

Parameter	Core	Listening	Reading	Writing	Speaking
Content knowledge	Unfamiliar	Mix of familiar and unfamiliar	Mix of familiar and unfamiliar	Familiar	Familiar
	There is a broad candidature so this is dealt with by selecting only clear topics accessible to the general reader. No expectation of knowledge of British culture. In some cases, domain or population specific version will include some level of expected knowledge of that domain or culture.				
Linguistic parameters					
Lexical range	The papers have as their basis the British Council/EAQUALS Core Inventory, which can be found at: http://www.teachingenglish.org.uk/article/british-council-eaquals-core-inventory-general-english-0 The language of the Aptis papers is carefully controlled, with clear specification of grammar and vocabulary for each task type (input and expected output) – lexical profiles are provided for all input texts (including instructions and prompts and are based on the Compleat Lexical Tutor (www.lextutor.ca)				
Structural range					
Functional range					

5.2.2.3. Text length

The amount of input/output depends on the paper, with texts of up to 750 words in the reading, expected outputs ranging from 30 seconds to two minutes in the speaking paper, and from 20 to 150 words and writing paper. The fact that the different tasks are designed to allow test takers at the different CEFR levels an opportunity to demonstrate their ability (thus 'biasing for best' – Swain, 1984) means that a variety of tasks reflecting different types of output are required. The Aptis papers offer a broad coverage of the constructs they are attempting to measure, thus facilitating claims as to their likely validity in predicting language ability across a range of levels.

5.2.2.4. Writer/speaker relationship

Setting up different relationships can impact on performance (see Porter and O'Sullivan, 1999). Therefore, where appropriate throughout the Aptis system, efforts have been made to specify (usually within the purpose of the task) the intended interlocutor or audience. An example of specifying different audiences is outlined in the discourse mode section above, where the test taker responds to a friend (informally) and then to a person of high status (formally) on the same topic.

5.2.2.5. Nature of information

Since more concrete topics/inputs are less difficult to respond to than more abstract ones and the intended candidature is generally at the intermediate or lower levels, the decision was made early in the development process to use texts that were more concrete in nature. Weir (2005, p. 74) argues that "Abstract information may in itself be cognitively as well as linguistically more complex and more difficult process". However, in the productive tasks, we have deliberately ensured that the expected output will vary from the concrete to the more abstract. An example of this is where three questions are asked (e.g. in relation to a photograph), the first will typically ask for a basic description, the second will ask the candidate to relate the input to their own experience, while the third will ask the candidate to speculate or offer an opinion. Thus, the cognitive load of the task gradually increases and, with it, the difficulty of the task.

5.2.2.6. Topic familiarity

Greater topic familiarity tends to result in superior performance. Therefore (for a similar reason to that outlined above), it was decided that topics would be used which were likely to be known to the candidates. Difficulty would be manipulated by increasing the cognitive load (through increasing or reducing parameters such as planning time, degree of concreteness, stipulation of audience and expected purpose, and amount of output).

5.2.3. The administrative parameters

The administrative parameters refer to conditions under which the test is taken. All major examination providers take great care to ensure that their tests are administered under similar and appropriate conditions for all candidates and that all aspects of the test process are secure. As can be seen in Table 5.4., the British Council ensures that these parameters are clearly outlined and reflected in all test administration practices by creating a series of guidelines for all those involved in test delivery. Delivery is also routinely monitored so that all procedures and regulations are fully followed. Even though Aptis is not seen as a high stakes test, the philosophy that drives all aspects of the test means that the British Council treats Aptis in the same way as it deals with the other high stakes tests it administers across the world.

Table 5.4: Administration-related evidence of the Aptis papers

Administrative parameters	
Physical conditions	Physical conditions for all tests are set out in the Administrator Guidelines, to which all delivering centres have access. Close monitoring of delivery is essential, particularly (though not exclusively) for the listening and speaking papers, where interference from nearby candidates is a risk. The British Council has extensive experience delivering tests across the world and does so for many language examination boards, as well as for general education examination boards and professional bodies.
Uniformity of administration	With the computer delivered versions of the papers, this is not a major issue, though there is a clear dependence on the physical conditions being appropriate. Administration is strictly controlled, and Aptis is treated by the British Council as any of the high stakes tests it administers around the world. With pen and paper versions, there are clearly set out procedures, which are monitored at the local level and on occasion from outside.
Security	This is less of an issue for Aptis than for major high stakes tests such as IELTS. However, security is seen as important and all test papers and test data are routinely stored (and securely destroyed when called for) in the same way we deal with high stakes tests. Security issues are dealt with in the Invigilator Guidelines and the Administrator Guidelines.

5.3. The scoring system

The key areas of scoring validity for the core, reading and listening papers include:

- accuracy of the answer key
- accuracy of data recording
- item performance
- internal consistency
- standard error of measurement (SEM).

Data from a series of trials held around the world formed the basis of the analysis results reported in Tables 5.5 and 5.6.

Table 5.5: Scoring system evidence – Aptis machine scored papers

Parameter	Core	Listening	Reading
Accuracy of the answer key	Answer keys are systematically checked on production of task/item, then again both pre and post test administration.		
Marker reliability	When taken online, responses are automatically scored within the system. This is the most accurate procedure. When a pen & paper version is taken, responses are manually input and routinely checked. However, Aptis is about to move to a Optical Mark Reader (OMR) to capture test scores – expected reliability is 99.98%.		
Item performance	All items are routinely trialled with a large (100+) representative sample of candidates from a range of countries. At this point, logit values (required in order to include items in the item bank) and other important data are collected (facility, point biserial, infit) in order to ensure that only properly functioning items are included in the test papers. Items are also routinely analysed post test delivery to confirm that they are working as expected.		
Internal consistency	0.98	0.91	0.95
SEM	4%	6%	7%

Table 5.6: Scoring validity of Aptis Speaking and Writing papers

Parameter	Writing	Speaking
Rating scale	The rating scales used for Aptis writing was developed based directly on the descriptors from the CEFR. The scales are task specific (one each for tasks 2, 3 and 4) and can be seen in Appendix 2.	The rating scales used for Aptis speaking was also based directly on the descriptors from the CEFR. The scale is test specific (a single scale is used for all four tasks) and can be seen in Appendix 4.
Rater selection	Minimum requirements for rater selection are set out in the Aptis guidelines. Experience in teaching and assessing at a range of levels is considered vital.	
Rater training	All raters are trained using materials based on the CEFR and representing the Aptis test tasks. Trainers are examiners who have received additional training.	
Rater monitoring	Raters are routinely monitored during to ensure they are on level. This is done in two ways: 1. Control scripts (pre-scored by expert raters) are fed to the rater during every session (approx. 5% of all performances marked), failure to meet pre-set conditions can result in removal from the system pending additional training. 2. Data from test scoring sessions are routinely analysed to ensure that all markers are on level.	
Rater agreement	0.94	0.91
Rater consistency	No inconsistent raters remain in the system, perhaps the attrition rate is due to the constant monitoring.	
Estimated SEM	7%	7%
Rating conditions	Raters may mark scripts in their own work environment, though they are given clear and strict instructions relating to the conduct of the assessment.	
Grading and awarding	Since all final decisions are made within the system, a unique approach to dealing with SEM. This is described in Section 5.3.1	

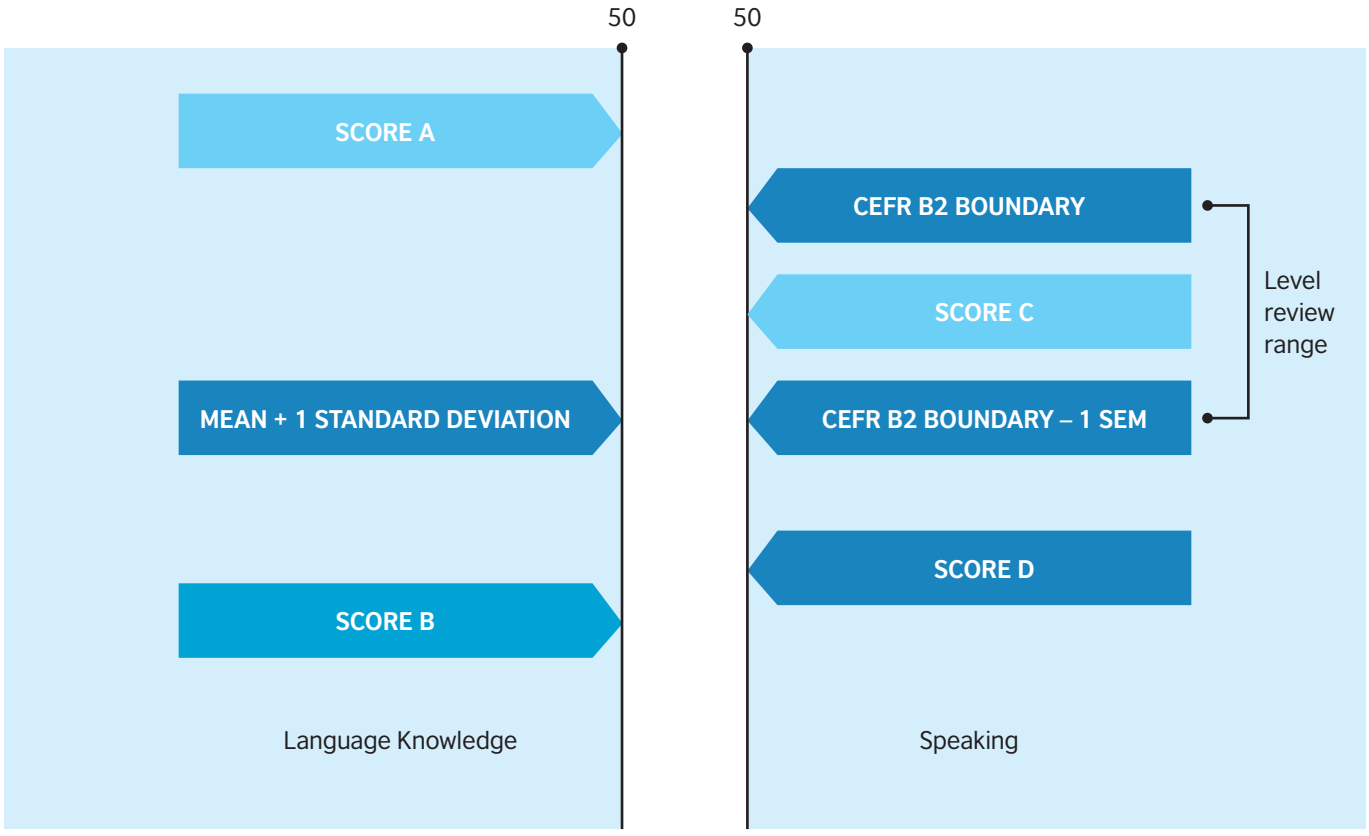
5.3.1. Using the core score to resolve boundary cases

The language knowledge score contributes to the overall CEFR level allocation in the following way.

Where a candidate achieves a score on their skills paper (in this example we are looking at speaking) that falls within 1 standard error of measurement (SEM) of a CEFR level boundary (e.g. achieving a score of 17 when the cut-score for B2 is 18), then their score on the language knowledge paper is taken into consideration when deciding whether they should remain at the lower level or be upgraded to the higher level. To receive this upgrade, they should perform significantly above the average (we set this at 1 standard deviation above the worldwide mean). This system greatly increases the accuracy of the CEFR level decisions and contributes significantly to the increased reliability of the outcomes.

In the example shown in Figure 5.1., a candidate who achieves score A on the language knowledge paper which is clearly above the review point (mean plus 1 standard deviation), will have their speaking score reviewed. If, like score C, it falls within the level review range (boundary point minus 1 SEM), then the person in this case will be awarded a B2 (rather than the lower B1). If it falls below this range (score D), then no action will be taken. If the candidate scores below the review point for language knowledge (score B), then no action is taken regarding the speaking paper score, regardless of where the speaking paper score lies in relation to the level review range.

Figure 5.1: Example of how the language knowledge score is used



5.3.2. Conclusions from this phase

The evidence presented here, from the overview of the test and its rationale and from the trials, strongly suggests a set of test papers that are working well to offer an accurate indication of an individual's language ability. The stability of the different papers, as shown in the tables, indicates that Aptis meets the expectations of a high stakes, international examination.

5.4. Claims

In keeping with the approach suggested in earlier linking projects (e.g. O'Sullivan, 2009), we now arrive at the stage when substantial claims regarding a test can be made.

Since Aptis is a completely new test, it was felt that the critical review stage suggested by O'Sullivan (2009) would not be needed, as the procedures outlined in this technical report act as a validation of the test. Of course, it would be naïve to think that the validation process ends with this report. On the contrary, this marks the formal beginning of the whole process. In recognition of the fact that test validation is an ongoing, long-term process, the British Council has undertaken two valuable initiatives. The first of these is the creation of the British Council Assessment Research Awards and Grants (the first of which were confirmed in early 2013). This initiative is designed to gather a broad range of validity evidence from external researchers across the world and is expected to contribute greatly to the test (in much the same way as the IELTS Joint Funded Research Scheme has for that test). The initiative is also designed to support young researchers with a series of small awards to help them complete work important to their careers. The other initiative is the revitalisation of the British Council's in-house research expertise. It is planned that this combination of internal and external research will add significantly to the validity evidence in support of various uses of Aptis in the coming years.

In order to lend support to claims of a link between the CEFR and the Aptis boundary points, we first completed the specification forms as suggested in the Council of Europe's Manual (2009). The evidence emerging from this activity supported the progression to the next phase, that of formal standard setting. The report of the standard-setting events presented here offer a strong vindication of the claims made by the British Council of the veracity of their CEFR-related claims. Finally, the validation stage demonstrated the test's accuracy and validity (in terms of content coverage and relevance; appropriacy of cognitive challenge; delivery and linguistic parameters; and scoring system).

All of this evidence combines to support both the validity of the test for use as a measure of general proficiency and the accuracy and appropriacy of the claimed links to the CEFR.

6.0 | SUMMARY AND DISCUSSION

The Aptis development project marked a new era for the British Council, even though it had been involved in a number of test development projects in the past, most notably ELTS (later IELTS). The decision was taken at an early stage in the project that the test should reflect best practice in the area of language testing and also 'fit' with the British Council's ambitions in the area of assessment literacy. These ambitions relate to the aim of offering world-class advice and consultancy to the many governments, institutions and corporation it works with across the globe. To make the most of the opportunities offered to the British Council itself and to its many partners in the UK and beyond, a wide-ranging assessment literacy agenda has been envisaged in which all British Council staff will be given the opportunity to learn about assessment. In addition, the plan is to pass on this knowledge and expertise to clients so that they can begin to make more informed decisions when it comes to assessment.

Aptis was developed as a low to medium stakes test to be used by large institutions such as education ministries, recruitment agencies and corporations in a variety of situations where an accurate, though affordable, estimation of the language levels of their employees or prospective employees was required.

The decision to undertake a formal CEFR linking project, normally the domain of high stakes tests, reflected a will to continue to push the boundaries of language testing.

The success of the project, as presented in this report, should not be taken as an end in itself. As already indicated, the British Council is committed to a long-term exploration of issues around the validation of Aptis and any future tests it is involved with.

6.1. Summary of the main findings

The project findings can be summarised as follows:

1 The Aptis papers offer a broad measure of ability across the different skills, as well as the key area of knowledge of the system of the language.
2. The Aptis test papers are robust in terms of quality of content and accuracy and consistency of decisions.
3. The CEFR boundary points suggested are robust and accurate.

6.2. Limitations

As with any project of this nature, there are limitations to this project. Pressure of time means that ongoing work to further support the psychometric qualities of the test cannot be included in this report, although this evidence will be made public in a future technical report.

6.3. Concluding comments

The project reported here was designed to offer evidence of the validity of claims of a link between the boundary points across the various Aptis skills papers and the CEFR. The fact that the project has provided evidence in support of these claims is of great importance to the British Council and the end-users of the test.

The development of Aptis and the completion of this project marks a significant beginning of the future of the British Council and high quality test development and validation.

REFERENCES

Abad Florescano, A., O'Sullivan, B., Sanchez Chavez, C., Ryan, D. E., Zamora Lara, E., Santana Martinez, L. A., Gonzalez Macias, M. I., Maxwell Hart, M., Grounds, P. E., Reidy Ryan, P., Dunne, R. A. and Romero Barradas, T. de E. (2011). Developing affordable 'local' tests: the EXAVER project. In Barry O'Sullivan (ed), *Language Testing: Theory & Practice (pp. 228-243).* Oxford: Palgrave Macmillan.

Alderson, J. C., Clapham, C & Wall, D. (1995). *Language Test Construction and Evaluation.* Cambridge: Cambridge University Press.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion referenced tests. *Review of Educational Research*, 56, pp. 137-172.

Berry, V. (2007). *Personality Differences and Oral Test Performance.* Frankfurt: Peter Lang.

Chizek, G. J. and Bunch, M. B. 2007. *Standard Setting.* Thousand Oaks, CA: Sage.

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

Council of Europe. (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment: Manual.* Strasburg: Council of Europe, Language Policy Division.

Council of Europe. (Undated). *Relating language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment – writing samples.* Retrieved January 10, 2013 from http://www.coe.int/t/dg4/education/elp/elp-reg/Source/Key_reference/exampleswriting_EN.pdf

Ginther, A. (2001). *Effects of the Presence and Absence of Visuals on Performance on TOEFL CBT Listening-Comprehensive Stimuli. TOEFL Research Report 66.* Princeton: Educational Testing Services.

Kane, M. T. (1992). An argument-based approach to validity, *Psychological Bulletin* 112 (3), pp. 527–535.

Kantarcıoğlu, E. (2012). *A Case-Study of the Process of Linking an Institutional English Language Proficiency Test (COPE) for Access to University Study in the Medium Of English to the Common European Framework for Languages: Learning, Teaching and Assessment.* (Unpublished PhD thesis.) University of Roehampton, London.

Kantarcıoğlu, E., Thomas, C., O'Dwyer, J. and O'Sullivan, B. (2010). The COPE linking project: a case study. In Waldemar Martyniuk (ed.) *Aligning Tests with the CEFR: Case studies and reflections on the use of the Council of Europe's Draft Manual* (pp. 102-118). Cambridge: Cambridge University Press

Khalifa, H. & Weir, C., J. (2009). *Examining Reading.* Cambridge: Cambridge University Press.

Maghera, D. & Rutherford, K. (2013). *Flexibility and Accessibility in a large-scale test of English proficiency.* Paper presented at the 3rd International Teacher Educator Conference, Hyderabad, India.

Messick, S. (1975). The standard program: Meaning and values in measurement and evaluation. *American Psychologist,* 30, pp. 955-966.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist,* 35, pp. 1012–1027.

Messick, S. (1989). Validity. In R. L. Linn (ed.), *Educational measurement (3rd edition)* London, NY: McMillan.

Mislevy R. J., Steinberg, L. S. & Almond, R. G. (2003). On the structure of educational assessments, *Measurement: Interdisciplinary Research and Perspectives* 1 (1), pp. 3-62.

Mislevy, R. J., Steinberg, L. S. & Almond, R. G. (2002). Design and analysis in task-based language assessment, *Language Testing* 19 (4), pp. 477–496.

O'Sullivan, B. (2002). Learner Acquaintanceship and Oral Proficiency Test Pair-Task Performance. *Language Testing,* 19 (3), pp. 277-295

O'Sullivan, B. (2005). *Levels Specification Project Report.* (Internal report.) Zayed University, United Arab Emirates.

O'Sullivan, B. (2008). *Modelling Performance in Tests of Spoken Language.* Frankfurt: Peter Lang.

O'Sullivan, B. (2009). *City & Guilds Communicator IESOL Examination (B2) CEFR Linking Project.* London: City & Guilds.

O'Sullivan, B. (2011). *Language Testing. In James Simpson (ed) Routledge Handbook of Applied Linguistics* (pp. 259-273). Oxford: Routledge.

O'Sullivan, B. (2015). *Aptis Test Development Approach.* Aptis Technical ReportTR/2015/003: British Council.

O'Sullivan, B. & Weir, C. (2011). Language Testing and Validation. In Barry O'Sullivan (ed) *Language Testing: Theory & Practice* (pp. 13-32). Oxford: Palgrave Macmillan.

Porter, D. & O'Sullivan, B. (1999). The effect of audience age on measured written performance. *System,* 27, pp. 65–77.

QALSPELL. 2004. *Quality Assurance in Language for Specific Purposes, Estonia, Latvia, Lithuania.* Leonardo da Vinci funded project. Website accessed June 8, 2008: http://www.qalspell.ttu.ee/

Swain, Merrill. (1984). Teaching and testing communicatively. *TESL Talk* Vol 15, No.s 1 and 2, pp. 7-18.

Unsworth, L. (2001). *Teaching multiliteracies across the curriculum: Changing contexts of text and image in classroom practice.* Buckingham: Open University Press.

Wagner, E. (2008). Video listening tests: What are they measuring? *Language Assessment Quarterly,* 5, pp. 218-243.

Weir, C. J. (2005). *Language Testing and Validation: an evidence-based approach.* Oxford: Palgrave.

Wu, J. R. W. & Wu, R. Y. F. (2010). Relating the GEPT Reading Comprehension Test to the CEFR. In Waldemar Martyniuk (ed.) *Aligning Tests with the CEFR: Case studies and reflections on the use of the Council of Europe's Draft Manual* (pp. 204-224). Cambridge: Cambridge University Press.

APPENDIX 1: COMPLETED SPECIFICATION FORMS

CEFR DRAFT LINKING MANUAL SPECIFICATION FORMS FOR APTIS

Completed September - December 2013

Section A2: Forms for Describing the Examination (Chapter 4)

GENERAL EXAMINATION DESCRIPTION			
1. General Information			
Name of examination	Aptis		
Language tested	English		
Examining institution	British Council		
Versions analysed (date)	August 2013		
Type of examination	☑ International ☐ National ☐ Regional ☐ nstitutional		
Purpose	To test general proficiency in English – four skills plus a language knowledge paper		
Target population	☐ Lower Sec ☑ Upper Sec ☑ Uni/College Students ☑ Adult		
No. of test takers per year	New test		
2. What is the overall aim?	To allow the test user to draw inferences, based on test performance, on the general language proficiency level of a test taker or population across the four skills.		
3. What are the more specific objectives?	If available describe the needs of the intended users on which this examination is based. Aptis is targeted specifically at large institutions worldwide who require some evidence of the language ability of their learners or employees. Aptis is not a certificated examination.		
4. What is/are principal domain(s)?	☑ Public ☑ Personal ☐ Occupational ☑ Educational		
5. Which communicative activities are tested?	☑ 1 Listening comprehension ☑ 2 Reading comprehension ☐ 3 Spoken interaction ☑ 4 Written interaction ☑ 5 Spoken production ☑ 6 Written production ☐ 7 Integrated skills ☐ 8 Spoken mediation of text ☐ 9 Written mediation of text ☑ 10 Language usage ☑ 11 Other: (specify): Written Interaction	*Name of Subtest(s)* 1. Core (Grammar & Vocabulary) 2. Reading 3. Listening 4. Writing 5. Speaking	Duration 25 minutes 40 minutes 25-50 minutes 40 minutes 15 minutes
6. What is the weighting of the different subtests in the global result?	Aptis is designed to offer a profile of language ability, with no 'overall' CEFR level reported. Where a client requires an overall grade, the ratio of importance of the skills is first agreed, then this ration is used as the basis of any calculation.		

7. Describe briefly the structure of each subtest	**Core**	**Listening**	**Reading**	**Writing**	**Speaking**
	Grammar – focus on grammatical form (including discourse usage) using MCQ items Vocabulary – focus on word definition, usage, synonyms, collocations	1. Listening for detail (pragmatic competence) – MCQ items 2. Listening for overall meaning – MCQ items 3. Listening for detail – note taking 4. Listening for detail – MCQ items	1. Careful local reading – MCQ cloze items 2. Careful global reading – re-building text 3. Global meaning – gapped text/ matching 4. Global reading – overall meaning	1. Form filling 2. Short extended guided writing (personal information) 3. Interactive writing (social media) semi-guided 4. Extended writing – informal and formal text	1. Personal information questions 2. Short answer non-personal questions (picture prompt) 3. Describe and compare questions (2 picture prompts) 4. Extended output based on prompt
8. What type(s) of responses are required?				**Subtests used in**	
	☑ Multiple-choice			Cg, L, R	
	☐ True/False				
	☑ Matching			Cv	
	☑ Ordering			R2	
	☑ Gap fill sentence			Cv	
	☑ Sentence completion			Cv;, Cg	
	☑ Gapped text / cloze, selected response			R1	
	☑ Open gapped text / cloze			R3	
	☑ Short answer to open question(s)			W1	
	☑ Extended answer (text)			W2, W3, W3	
	☐ Interaction with examiner				
	☐ Interaction with peers				
	☑ Other (Short answer - spoken)			S1, S2	
	☑ Other (Extended answer - spoken)			S3, S4	
9. What information is published for candidates and teachers?	☑ Overall aim ☑ Principal domain(s) ☑ Test subtests ☑ Test tasks ☑ Sample test papers ☐ Video of format of oral			☑ Sample answer papers ☑ Marking schemes ☐ Grading schemes	

10. Where is this accessible?		**Subtests used in**
	☑ On the website ☐ From bookshops ☑ In test centres ☑ On request from the institution ☐ Other	
11. What is reported?	☐ Global grade ☑ Grade per subtest (scale 0-50) ☑ CEFR Profile	☐ Global grade plus graphic profile ☐ Profile per subtest

Form A1: General Examination Description (continued)

Test development	Short description and/or references
1. What organisation decided that the examination was required?	☑ Own organisation/school ☐ A cultural institute ☐ Ministry of Education ☐ Ministry of Justice ☐ Other: specify: _____
2. If an external organisation is involved, what influence do they have on design and development?	Not Applicable
3. If no external organisation was involved, what other factors determined design and development of examination?	☑ A needs analysis ☑ Internal description of examination aims ☑ Internal description of language level Based on CEFR ☑ A syllabus or curriculum ☑ Profile of candidates
4. In producing test tasks are specific features of candidates taken into account? *Every effort is taken to ensure that the questions included are free from bias and are in line with the guidelines of: APA (American Psychological Association) standards of educational and psychological testing, AERA, NAPA and NCME.*	☑ Linguistic background (L1) ☑ Language learning background ☑ Age ☑ Educational level ☑ Socio-economic background ☑ Social-cultural factors ☑ Ethnic background ☑ Gender
5. Who writes the items or develops the test tasks?	A specially trained team of British Council experienced teachers (up to 250 hours of training in assessment & item design and writing)

Test development	Short description and/or references
6. Have test writers guidance to ensure quality?	☑ Training ☑ Guidelines ☑ Checklists ☑ Examples of valid, reliable, appropriate tasks: ☑ Calibrated to CEFR level description ☐ Calibrated to other level description: _____
7. Is training for test writers provided?	☑ Yes ☐ No
8. Are test tasks discussed before use?	☑ Yes ☐ No
9. If yes, by whom?	☑ Individual colleagues ☑ Internal group discussion ☐ External examination committee ☑ Internal stakeholders ☐ External stakeholders
10. Are test tasks pretested?	☑ Yes ☐ No
11. If yes, how?	With a population of over 100 candidates who have been identified by centres as being at the appropriate level.
12. If no, why not?	
13. Is the reliability of the test estimated?	☑ Yes ☐ No
14. If yes, how?	☑ Data collection and psychometric procedures – Cronbach's alpha ☑ Other: Scorer reliability (inter- & intra-rater) is calculated for the rating of the writing tasks
15. Are different aspects of validity estimated?	☐ Face validity – during piloting - questionnaires to teachers in examination centres ☑ Test taker related ☑ Test task related ☑ Scoring system related
16. If yes, describe how.	By a team of trained experts during the development process, analysis as part of routine quality assurance procedures

Form A2: Test Development (continued)

Marking: Subtest	Listening
1. How are the test tasks marked?	For receptive test tasks: ☑ Optical mark reader (coming in 2013) ☑ Clerical marking
2. Where are the test tasks marked?	☑ Centrally (on computer versions) ☑ Locally: ☑ By local teams ☑ By individual examiners
3. What criteria are used to select markers?	Qualified teachers Experienced teachers
4. How is accuracy of marking promoted?	☑ Regular checks by co-ordinator ☑ Training of markers ☐ Moderating sessions to standardise judgements ☐ Using standardised examples of test tasks: ☐ Calibrated to CEF –levels ☐ Not calibrated to CEF or other description
5. Describe the specifications of the rating criteria of productive and/or integrative test tasks.	N/A
6. Are productive or integrated test tasks single or double rated?	N/A
7. If double rated, what procedures are used when differences between raters occur?	N/A
8. Is inter-rater agreement calculated?	N/A
9. Is intra-rater agreement calculated?	N/A

Marking: Subtest	Reading
1. How are the test tasks marked?	For receptive test tasks: ☑ Optical mark reader (coming in 2013) ☑ Clerical marking
2. Where are the test tasks marked?	☑ Centrally (on computer versions) ☑ Locally: ☑ By local teams ☑ By individual examiners
3. What criteria are used to select markers?	Qualified teachers Experienced teachers
4. How is accuracy of marking promoted?	☑ Regular checks by co-ordinator ☑ Training of markers ☐ Moderating sessions to standardise judgements ☐ Using standardised examples of test tasks: ☐ Calibrated to CEF –levels ☐ Not calibrated to CEF or other description
5. Describe the specifications of the rating criteria of productive and/or integrative test tasks.	N/A
6. Are productive or integrated test tasks single or double rated?	N/A
7. If double rated, what procedures are used when differences between raters occur?	N/A
8. Is inter-rater agreement calculated?	N/A
9. Is intra-rater agreement calculated?	N/A

Marking: Subtest	Writing
1. How are the test tasks marked?	For receptive test tasks (editing): ☐ Optical mark reader ☐ Clerical marking For productive or integrated test tasks: ☑ Trained examiners ☐ Teachers
2. Where are the test tasks marked?	☑ Centrally ☐ Locally: ☐ By local teams ☐ By individual examiners
3. What criteria are used to select markers?	Qualified teachers Experienced teachers Must have passed an accreditation test
4. How is accuracy of marking promoted?	☑ Regular checks by co-ordinator ☑ Training of markers/raters ☑ Moderating sessions to standardise judgements ☑ Using standardised examples of test tasks: ☑ Calibrated to CEF –levels ☐ Not calibrated to CEF or other description
5. Describe the specifications of the rating criteria of productive and/or integrative test tasks.	☑ One holistic score for each task ☐ Marks for different aspects for each task ☐ Rating scale for overall performance in test ☐ Rating grid for aspects of test performance ☑ Rating scale for each task ☐ Rating grid for aspects for each task ☐ Rating scale bands are defined, but not to CEFR ☑ Rating scale bands are defined in relation to CEFR
6. Are productive or integrated test tasks single or double rated?	☐ Single rater ☐ Two simultaneous raters ☑ Double marking of scripts (random) ☑ Other: specify: Approx. 5% of all scripts are pre-marked by multiple expert raters. Failure to mark to standard results in withdrawal until additional training is passed

Marking: Subtest	Writing
7. If double rated, what procedures are used when differences between raters occur?	☑ Use of third rater and that score holds ☐ Use of third marker and two closest marks used ☐ Average of two marks ☐ Two markers discuss and reach agreement
8. Is inter-rater agreement calculated?	☑ Yes measuring the inter-rater reliability currently with Spearman Rho & multi-faceted Rasch analysis ☐ No
9. Is intra-rater agreement calculated?	☑Yes – multi-faceted Rasch analysis used ☐ No

Marking: Subtest	Speaking
1. How are the test tasks marked?	For receptive test tasks (editing): ☐ Optical mark reader ☐ Clerical marking For productive or integrated test tasks: ☑ Trained examiners ☐ Teachers
2. Where are the test tasks marked?	☑ Centrally ☐ Locally: ☐ By local teams ☐ By individual examiners
3. What criteria are used to select markers?	Qualified teachers Experienced teachers Must have passed an accreditation test
4. How is accuracy of marking promoted?	☑ Regular checks by co-ordinator ☑ Training of markers/raters ☑ Moderating sessions to standardise judgements ☑ Using standardised examples of test tasks: ☑ Calibrated to CEF –levels ☑ Not calibrated to CEF or other description

Marking: Subtest	Speaking
5. Describe the specifications of the rating criteria of productive and/or integrative test tasks.	☑ One holistic score for each task ☐ Marks for different aspects for each task ☐ Rating scale for overall performance in test ☐ Rating grid for aspects of test performance ☑ Rating scale for each task ☐ Rating grid for aspects for each task ☐ Rating scale bands are defined, but not to CEFR ☑ Rating scale bands are defined in relation to CEFR
6. Are productive or integrated test tasks single or double rated?	☐ Single rater ☐ Two simultaneous raters ☑ Double marking of scripts (random) ☑ Other: specify: Approx. 5% of all scripts are pre-marked by multiple expert raters. Failure to mark to standard results in withdrawal until additional training passed
7. If double rated, what procedures are used when differences between raters occur?	☑ Use of third rater and that score holds ☐ Use of third marker and two closest marks used ☐ Average of two marks ☐ Two markers discuss and reach agreement
8. Is inter-rater agreement calculated?	☑ Yes measuring the inter-rater reliability currently with Spearman Rho & multi-faceted Rasch analysis ☐ No
9. Is intra-rater agreement calculated?	☐ Yes – multi-faceted Rasch analysis used ☐ No

Form A3: Marking

Grading: Listening	Complete a copy of this form for each subtest. Short description and/or reference
1. Are pass marks and/or grades given?	☐ Pass marks ☑ CEFR levels ☑ Scale scores (0-50)
2. Describe the procedures used to establish pass marks and/or grades and cut-scores	The boundaries are set using a modified Angoff standard-setting procedure, described in full in the standard-setting section of this report.
3. If only pass/fail is reported, how are the cut-off scores for pass/fail set?	N/A
4. If grades are given, how are the grade boundaries decided?	See item 2 in this table
5. How is consistency in these standards maintained?	Consistency is maintained by ensuring that the parallel versions of the test are equivalent. Tests are compiled from an item bank and must reflect a specified difficulty profile (IRT-based). In addition, item writers are carefully trained and follow the specifications, while a quality assurance system pre-proofs all items prior to pilot testing.

Grading: Reading	Complete a copy of this form for each subtest. Short description and/or reference
1. Are pass marks and/or grades given?	☐ Pass marks ☑ CEFR levels ☑ Scale scores (0-50)
2. Describe the procedures used to establish pass marks and/or grades and cut-scores	The boundaries are set using a modified Angoff standard-setting procedure, described in full in the standard-setting section of this report.
3. If only pass/fail is reported, how are the cut-off scores for pass/fail set?	N/A
4. If grades are given, how are the grade boundaries decided?	See item 2 in this table
5. How is consistency in these standards maintained?	Consistency is maintained by ensuring that the parallel versions of the test are equivalent. Tests are compiled from an item bank and must reflect a specified difficulty profile (IRT-based). In addition, item writers are carefully trained and follow the specifications, while a quality assurance system pre-proofs all items prior to pilot testing.

Grading: Writing	Complete a copy of this form for each subtest. Short description and/or reference
1. Are pass marks and/or grades given?	☐ Pass marks ☑ CEFR levels ☑ Scale scores (0-50)
2. Describe the procedures used to establish pass marks and/or grades and cut-scores	The boundaries are set using a modified Angoff standard setting procedure and is described in full in the Standard Setting section of this report.
3. If only pass/fail is reported, how are the cut-off scores for pass/fail set?	N/A
4. If grades are given, how are the grade boundaries decided?	See item 2 in this table
5. How is consistency in these standards maintained?	Item writers are carefully trained and follow the specifications, while a quality assurance system pre-proofs all items prior to pilot testing. Test versions are routinely analysed for consistency of level using multi-faceted Rasch analysis.

Grading: Speaking	Complete a copy of this form for each subtest. Short description and/or reference
1. Are pass marks and/or grades given?	☐ Pass marks ☑ CEFR levels ☑ Scale scores (0-50)
2. Describe the procedures used to establish pass marks and/or grades and cut-scores	The boundaries are set using a modified Angoff standard-setting procedure, described in full in the standard-setting section of this report.
3. If only pass/fail is reported, how are the cut-off scores for pass/fail set?	N/A
4. If grades are given, how are the grade boundaries decided?	See item 2 in this table
5. How is consistency in these standards maintained?	Item writers are carefully trained and follow the specifications, while a quality assurance system pre-proofs all items prior to pilot testing. Test versions are routinely analysed for consistency of level using multi-faceted Rasch analysis.

Form A4: Grading

Results	Short description and/or reference
1. What results are reported to candidates?	☐ Global grade or pass / fail ☑ CEFR level ☐ Global grade plus profile across subtests ☑ Scaled score (0-50)
2. In what form are results reported?	☐ Undefined grades (e.g. "C") ☑ Level on a defined scale ☑ Diagnostic profiles
3. On what document are results reported?	☐ Letter or email ☑ Report form to candidate ☐ Certificate / Diploma ☑ Online to client
4. Is information provided to help candidates to interpret results? Give details.	Details on the report form and on the dedicated website on the meaning of the CEFR levels that are used on the certificate – based on 'Can Do' statements.
5. Do candidates have the right to see the corrected and scored examination papers?	No
6. Do candidates have the right to ask for remarking?	Yes

Form A5: Reporting Results

Data analysis	Short description and/or reference
1. Is feedback gathered on the examinations?	☑ Yes, in the course of pre-testing and live testing ☐ No
2. If yes, by whom?	☑ Internal experts (colleagues) ☑ External experts ☐ Local examination institutes ☑ Test administrators ☑ Teachers ☑ Candidates
3. Is the feedback incorporated in revised versions of the examinations?	☑ Yes ☐ No
4. Is data collected to do analysis on the tests?	☑ On all tests ☐ On a sample of test takers: How large?: _____. How often?:_____ ☐ No
5. If yes, indicate how data are collected?	☑ During pre-testing ☑ During live examinations ☐ After live examinations
6. For which features is analysis on the data gathered carried out?	☑ Difficulty ☑ Discrimination ☑ Reliability ☑ Validity (content)
7. State which analytic methods have been used (e.g. in terms of psychometric procedures).	• Descriptive stats – measures of central tendency and dispersion • Classical item statistics • IRT – item level difficulty and item misfit • Qualitative feedback (how it works/rater remarks) • Inter-subtest correlations
8. Are performances of candidates from different groups analysed? If so, describe how.	Yes – bias analysis / DIF based on the candidate data performed during annual test review.
9. Describe the procedures to protect the confidentiality of data.	All scripts are handled and stored within secure areas. Data are analysed using spreadsheets held on a secure network drive. There is limited access to this data.
10. Are relevant measurement concepts explained for test users? If so, describe how.	Summary of how final scores are calculated is available.

Form A6: Data Analysis

Rationale for decisions (and revisions)	Short description and/or reference
Give the rationale for the decisions that have been made in relation to the examination or the test tasks in question.	Basic underlying philosophy is flexibility and accessibility For this reason: • Different delivery options – computer (tablets and iPad in 2013), phone, pen & paper • Client decides on which skills to test • Customisation of content possible
Is there a review cycle for the examination? (How often? Who by? Procedures for revising decisions)	No fixed time, though already (6 months after launch) we are returning to the Listening paper with a view to future amendments. It is part of the philosophy of the test, to constantly improve.

Form A7: Rationale for Decisions

Initial estimation of overall CEFR level
Short rationale, reference to documentation Aptis is not designed to offer a measure of ability at a single level, instead it measures across levels A1 to C – no attempt is made to distinguish between C1 and C2 (though this may be done in future iterations of the test).

Form A8: Initial Estimation of Overall Examination Level

Section A3: Specification: Communicative Language Activities (Chapter 4)

A3.1 Reception

Listening Comprehension	Short description and/or reference
1. In what contexts (domains, situations, ...) are the test takers to show ability?	Personal, Public and Educational
2. Which communication themes are the test takers expected to be able to handle?	As Aptis tests across the levels, different themes are found in different items, examples include: Self and Family, House and Home, Environment, Daily Life, Free time, Entertainment, Travel, Shopping, Food and Drink, Public Services, Places, Language, Time, Numbers, Weather, Measures and Shapes.
3. Which communicative tasks, activities and strategies are the test takers expected to be able to handle?	• Listening for detail in announcements and messages • Listen for speaker intent/mood/attitude
4. What text-types and what length of text are the test takers expected to be able to handle?	• Interpersonal dialogues and conversations • Broadcasts • Discussions • Instructions and directions • Telephone conversations
5. After reading the scale for Overall Listening Comprehension, given below, indicate and justify at which level(s) of the scale the subtest should be situated.	**Levels: A1 to C** **Justification (incl. reference to documentation)** The items, themes and foci of the input texts were drawn from the CEFR. Trialling of items indicates a broad range of difficulty Standard setting acts to triangulate this evidence

Form A9: Listening Comprehension

	Overall listening comprehension
C2	*Has no difficulty in understanding any kind of spoken language, whether live or broadcast, delivered at fast native speed.*
C1	*Can understand enough to follow extended speech on abstract and complex topics beyond his/her own field, though he/she may need to confirm occasional details, especially if the accent is unfamiliar.* *Can recognise a wide range of idiomatic expressions and colloquialisms, appreciating register shifts.* *Can follow extended speech even when it is not clearly structured and when relationships are only implied and not signalled explicitly.*
B2	*Can understand standard spoken language, live or broadcast, on both familiar and unfamiliar topics normally encountered in personal, social, academic or vocational life. Only extreme background noise, inadequate discourse structure and/or idiomatic usage influences the ability to understand.*
	Can understand the main ideas of propositionally and linguistically complex speech on both concrete and abstract topics delivered in a standard dialect, including technical discussions in his/her field of specialisation. *Can follow extended speech and complex lines of argument provided the topic is reasonably familiar, and the direction of the talk is sign-posted by explicit markers.*
B1	*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.*
	Can understand the main points of clear standard speech on familiar matters regularly encountered in work, school, leisure etc., including short narratives.
A2	*Can understand enough to be able to meet needs of a concrete type provided speech is clearly and slowly articulated.*
	Can 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.
A1	*Can follow speech which is very slow and carefully articulated, with long pauses for him/her to assimilate meaning.*

Relevant Subscales for Listening Comprehension	English
➢ Understanding conversation between native speakers	Page 66
➢ Listening as a member of an audience	Page 67
➢ Listening to announcements and instructions	Page 67
➢ Listening to audio media and recordings	Page 68
➢ Identifying cues and inferring	Page 72

Reading Comprehension	Short description and/or reference
1. In what contexts (domains, situations, ...) are the test takers to show ability?	Public, personal and educational
2. Which communication themes are the test takers expected to be able to handle?	As Aptis tests across the levels, different themes are found in different items, examples include: Self and Family, House and Home, Environment, Daily Life, Free time, Entertainment, Travel, Shopping, Food and Drink, Public Services, Places, Language, Time, Numbers, Weather, Measures and Shapes.
3. Which communicative tasks, activities and strategies are the test takers expected to be able to handle?	These are as outlined in the British Council/EAQUALS Core Inventory **Tasks:** • Completing texts (variety of text types) by inserting missing sentences / words into phrases • Completing short texts • Locating specific information **Activities:** • Reading for global comprehension • Reading for local detail **The language user may read:** • for gist • for specific information • for detailed understanding **Strategies:** • Planning: framing • Execution: Identifying cues and inferring from them • Evaluation: hypothesis testing • Repair: revising hypothesis
4. What text-types and what length of text are the test takers expected to be able to handle?	**Text types:** • Narratives • Explanations • Descriptions Text length: max 750 words (exc. Items)
5. After reading the scale for Overall Reading Comprehension, given below, indicate and justify at which level(s) of the scale the subtest should be situated.	**Levels: A1 to C** **Justification (incl. reference to documentation)** The items, themes and foci of the input texts were drawn from the CEFR. Trialling of items indicates a broad range of difficulty Standard setting acts to triangulate this evidence Test based on extensive research during development – using Khalifa & Weir's (2009) reading model

Form A10: Reading Comprehension

	Overall reading comprehension
C2	*Can understand and interpret critically virtually all forms of the written language including abstract, structurally complex, or highly colloquial literary and non-literary writings.* *Can understand a wide range of long and complex texts, appreciating subtle distinctions of style and implicit as well as explicit meaning.*
C1	*Can understand in detail lengthy, complex texts, whether or not they relate to his/her own area of speciality, provided he/she can reread difficult sections.*
B2	*Can read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively. Has a broad active reading vocabulary, but may experience some difficulty with low-frequency idioms.*
B1	*Can read straightforward factual texts on subjects related to his/her field and interest with a satisfactory level of comprehension.*
A2	*Can understand short, simple texts on familiar matters of a concrete type which consist of high frequency everyday or job-related language*
	Can understand short, simple texts containing the highest frequency vocabulary, including a proportion of shared international vocabulary items.
A1	*Can understand very short, simple texts a single phrase at a time, picking up familiar names, words and basic phrases and rereading as required.*

Relevant Subscales for Reading Comprehension	English
➢ Reading for orientation	Page 70
➢ Reading for information and argument	Page 70
➢ Identifying cues and inferring	Page 72

A3.2 Interaction

Written Interaction	Short description and/or reference
1. In what contexts (domains, situations, ...) are the test takers to show ability?	Personal, public and educational
2. Which communication themes are the test takers expected to be able to handle?	As Aptis tests across the levels, different themes are found in different items. Interactive writing is examined in Tasks 3 and 4 only. Examples of themes include: Environment, Entertainment, Travel, Shopping, Food and Drink, Public Services, Clubs
3. Which communicative tasks, activities and strategies are the test takers expected to be able to handle?	Tasks: (awareness of audience is the key in both tasks) • Social media reading and responding • Formal and informal writing on same topic **Strategies** • Planning • Execution • Evaluation • Repair
4. What kind of texts and text-types are the test takers expected to be able to handle?	• Social Media • Email
5. After reading the scale for Overall Written Interaction, given below, indicate and justify at which level(s) of the scale the subtest should be situated.	**Levels: B1 to C** **Justification (incl. reference to documentation)** The tasks, themes and foci of the input texts were drawn from the CEFR Tasks designed to elicit language at B1 and above Standard setting acts to triangulate this evidence

	Overall written interaction
C2	*As C1*
C1	*Can express him/herself with clarity and precision, relating to the addressee flexibly and effectively.*
B2	*Can express news and views effectively in writing, and relate to those of others.*
B1	*Can convey information and ideas on abstract as well as concrete topics, check information and ask about or explain problems with reasonable precision.*
	Can write personal letters and notes asking for or conveying simple information of immediate relevance, getting across the point he/she feels to be important.
A2	*Can write short, simple formulaic notes relating to matters in areas of immediate need.*
A1	*Can ask for or pass on personal details in written form.*

Relevant Subscales for Written Interaction	English
➢ Correspondence	Page 83
➢ Vocabulary Range & Control	Page 112
➢ Grammatical Accuracy	Page 114
➢ Sociolinguistic Appropriateness	Page 122
➢ Flexibility	Page 124
➢ Thematic Development	Page 125
➢ Coherence & Cohesion	Page 125

A3.3 Production

Written Production	Short description and/or reference
1. In what contexts (domains, situations, ...) are the test takers to show ability?	Home, office, place of study
2. Which communication themes are the test takers expected to be able to handle?	As Aptis tests across the levels, different themes are found in different items. Interactive writing is examined in Tasks 1 and 2 only. Examples of themes include: House and Home, Daily life, Free time, entertainment, Personal Information
3. Which communicative tasks, activities and strategies are the test takers expected to be able to handle?	Form filling (basic) Form filling (extended response)
4. What kind of texts and text-types are the test takers expected to be able to handle? The lists in CEFR 4.6.2 and 4.6.3 might be of help as a reference.	Descriptive Narrative Expository
5. After reading the scale for Overall Written Production, given below, indicate and justify at which level(s) of the scale the subtest should be situated. The subscales for written production in CEFR 4.4.1.2 listed after the scale might be of help as a reference.	**Levels: A1 to A2** Justification (incl. reference to documentation) The tasks, themes and foci of the input texts were d rawn from the CEFR Tasks designed to elicit language at level A Standard setting acts to triangulate this evidence

Form A13: Written Production

	Overall written interaction
C2	*Can write clear, smoothly flowing, complex texts in an appropriate and effective style and a logical structure which helps the reader to find significant points.*
C1	*Can write clear, well-structured texts of complex subjects, underlining the relevant salient issues, expanding and supporting of view at some length with subsidiary point, reasons and relevant examples, and rounding off with an appropriate conclusion.*
B2	Can write clear, detailed texts on a variety of subjects related to his/her field of interest, synthesising and evaluating information and arguments from a number of sources.
B1	*Can write straightforward connected texts on a range of familiar subjects within his/her field of interest, by linking a series of shorter discrete elements into a linear sequence.*
A2	*Can write a series of simple phrases and sentences linked with simple connecters like "and, "but" and "because".*
A1	*Can write simple isolated phrases and sentences.*

Relevant Subscales for Written Interaction	English
➢ Vocabulary Range & Control	Page 112
➢ Grammatical Accuracy	Page 114
➢ Notes, messages and forms	Page 84

Spoken Production	Short description and/or reference
1. In what contexts (domains, situations, ...) are the test takers to show ability?	Home, office, place of study
2. Which communication themes are the test takers expected to be able to handle?	As Aptis tests across the levels, different themes are found in different items. Interactive writing is examined in Tasks 1 and 2 only. Examples of themes include: Self and Family, House and Home, Environment, Daily Life, Free time, Entertainment, Travel, Shopping, Food and Drink, Public Services, Places, Language, Time, Numbers, Weather.
3. Which communicative tasks, activities and strategies are the test takers expected to be able to handle?	Responding to questions Describing Comparing Speculating
4. What kind of texts and text-types are the test takers expected to be able to handle?	Descriptive Narrative Expository
5. After reading the scale for Overall Written Production, given below, indicate and justify at which level(s) of the scale the subtest should be situated.	**Levels: A1 to C** **Justification (incl. reference to documentation)** The items, themes and foci of the input texts were drawn from the CEFR. Quality control of items indicates a broad range of difficulty Standard setting acts to triangulate this evidence

Form A14: Spoken Production

	Overall written production
C2	*Can write clear, smoothly flowing, complex texts in an appropriate and effective style and a logical structure which helps the reader to find significant points.*
C1	*Can write clear, well-structured texts of complex subjects, underlining the relevant salient issues, expanding and supporting of view at some length with subsidiary point, reasons and relevant examples, and rounding off with an appropriate conclusion.*
B2	Can write clear, detailed texts on a variety of subjects related to his/her field of interest, synthesising and evaluating information and arguments from a number of sources.
B1	*Can write straightforward connected texts on a range of familiar subjects within his/her field of interest, by linking a series of shorter discrete elements into a linear sequence.*
A2	*Can write a series of simple phrases and sentences linked with simple connecters like "and, "but" and "because".*
A1	*Can write simple isolated phrases and sentences.*

Relevant Subscales for Spoken Production	English
➢ Information Exchange (Task 1)	Page 81
➢ Being Interviewed (Task 1)	Page 82
➢ Vocabulary Range & Control	Page 112
➢ Grammatical Accuracy	Page 114
➢ Phonological Control	Page 117

Section A4: Specification: Communicative Language Competence (Chapter 4)

Forms concerning competence are again provided in the following order:

1. Reception
2. Interaction
3. Production
4. Mediation

A4.1 Reception

Those CEFR scales most relevant to Receptive skills have been used to create Table A3, which can be referred to in this section. Table A3 does not include any descriptors for "plus levels". The original scales consulted, some of which do define plus levels, include:

Linguistic Competence

- General Linguistic Range — English: page 110
- Vocabulary Range — English: page 112

Socio-linguistic Competence

- Socio-linguistic Appropriateness — English: page 122

Pragmatic Competence

- Thematic Development — English: page 125
- Cohesion and Coherence — English: page 125
- Propositional Precision — English: page 129

Strategic Competence

- Identifying Cues/Inferring — English: page 72

Linguistic Competence	Short description and/or reference
1. What is the range of lexical and grammatical competence that the test takers are expected to be able to handle?	This is clearly set out in the British Council/EAQUALS Core Inventory.
2. After reading the scale for Linguistic Competence in Table A3, indicate and justify at which level(s) of the scale the examination should be situated.	**Levels A1-C** **Justification (incl. reference to documentation)** • The items, themes and foci of the input texts were drawn from the CEFR, while the areas of linguistic competence are based on a Core Inventory which itself is very much driven by the CEFR. • Quality control of items ensures that the item and task writers continue to meet the expectations of the specifications. • The specifications are created in such a way as to encourage interaction between the item writers and the quality assurance team. • Standard setting acts to triangulate this evidence • Standard setting acts to triangulate this evidence • See Table A3

Socio-linguistic Competence	Short description and/or reference
3. What are the socio-linguistic competences that the test takers are expected to be able to handle: linguistic markers, politeness conventions, register, adequacy, dialect/accent, etc.?	These are clearly set out in the British Council/EAQUALS Core Inventory.
4. After reading the scale for Socio-linguistic Competence in Table A3, indicate and justify at which level(s) of the scale the examination should be situated.	**Levels A1-C** **Justification (incl. reference to documentation)** • The items, themes and foci of the input texts were drawn from the CEFR, while the areas of linguistic competence are based on a Core Inventory which itself is very much driven by the CEFR. • Quality control of items ensures that the item and task writers continue to meet the expectations of the specifications. • The specifications are created in such a way as to encourage interaction between the item writers and the quality assurance team. • Standard setting acts to triangulate this evidence • See Table A3

Form A19: Aspects of Language Competence in Reception (part)

TABLE A3: RELEVANT QUALITATIVE FACTORS FOR RECEPTION – appropriate cells are shaded				
	LINGUISTIC: Edited from General Linguistic Range; Vocabulary Range	**SOCIO-LINGUISTIC: Edited from Socio-linguistic Appropriateness**	**PRAGMATIC: Edited from Thematic Development and Propositional Precision**	**STRATEGIC: Identifying Cues and Inferring**
C2	*Can understand a very wide range of language precisely, appreciating emphasis and, differentiation. No signs of comprehension problems.* *Has a good command of a very broad lexical repertoire including idiomatic expressions and colloquialisms; shows awareness of connotative levels of meaning.*	*Has a good command of idiomatic expressions and colloquialisms with awareness of connotative levels of meaning.* *Appreciates fully the socio-linguistic and sociocultural implications of language used by native speakers and can react accordingly.*	*Can understand precisely finer shades of meaning conveyed by a wide range of qualifying devices (e.g. adverbs expressing degree, clauses expressing limitations).* *Can understand emphasis and differentiation without ambiguity.*	*As C1.*
C1	*Has a good command of a broad lexical repertoire.* *Good command of idiomatic expressions and colloquialisms.*	*Can recognise a wide range of idiomatic expressions and colloquialisms, appreciating register shifts; may, however, need to confirm occasional details, especially if the accent is unfamiliar.* *Can follow films employing a considerable degree of slang and idiomatic usage.* *Can understand language effectively for social purposes, including emotional, allusive and joking usage.*	*Can understand elaborate descriptions and narratives, recognising sub-themes, and points of emphasis.* *Can understand precisely the qualifications in opinions and statements that relate to degrees of, for example, certainty/uncertainty, belief/doubt, likelihood etc.*	*Is skilled at using contextual, grammatical and lexical cues to infer attitude, mood and intentions and anticipate what will come next.*
B2	*Has a sufficient range of language to be able to understand descriptions, viewpoints and arguments on most topics pertinent to his everyday life such as family, hobbies and interests, work, travel, and current events.*	*Can with some effort keep up with fast and colloquial discussions.*	*Can understand description or narrative, identifying main points from relevant supporting detail and examples.* *Can understand detailed information reliably.*	*Can use a variety of strategies to achieve comprehension, including listening for main points; checking comprehension by using contextual clues.*
B1	*Has enough language to get by, with sufficient vocabulary to understand most texts on topics such as family, hobbies and interests, work, travel, and current events.*	*Can respond to a wide range of language functions, using their most common exponents in a neutral register.* *Can recognise salient politeness conventions.* *Is aware of, and looks out for signs of, the most significant differences between the customs, usages, attitudes, values and beliefs prevalent in the community concerned and those of his or her own.*	*Can reasonably accurately understand a straightforward narrative or description that is a linear sequence of points.*	*Can identify unfamiliar words from the context on topics related to his/her field and interests.* *Can extrapolate the meaning of occasional unknown words from the context and deduce sentence meaning provided the topic discussed is familiar.*
A2	*Has a sufficient vocabulary for coping with everyday situations with predictable content and simple survival needs.*	*Can handle very short social exchanges, using everyday polite forms of greeting and address. Can make and respond to invitations, apologies etc.*	*Can understand a simple story or description that is a list of points.* *Can understand a simple and direct exchange of limited information on familiar and routine matters.*	*Can use an idea of the overall meaning of short texts and utterances on everyday topics of a concrete type to derive the probable meaning of unknown words from the context.*
A1	*Has a very basic range of simple expressions about personal details and needs of a concrete type.*	*Can understand the simplest everyday polite forms of: greetings and farewells; introductions; saying please, thank you, sorry etc.*	*No descriptor available.*	*No descriptor available.*

Pragmatic Competence	Short description and/or reference
5. What are the pragmatic competences that the test takers are expected to be able to handle: discourse competences, functional competences?	This is clearly set out in the British Council/ EAQUALS Core Inventory.
6. After reading the scale for Pragmatic Competence in Table A3, indicate and justify at which level(s) of the scale the examination should be situated.	**Level A1-C** **Justification (incl. reference to documentation)** The items, themes and foci of the input texts were drawn from the CEFR, while the areas of linguistic competence are based on a Core Inventory which itself is very much driven by the CEFR. Quality control of items ensures that the item and task writers continue to meet the expectations of the specifications. The specifications are created in such a way as to encourage interaction between the item writers and the quality assurance team. Standard setting acts to triangulate this evidence.
Strategic Competence	**Short description and/or reference**
7. What are the strategic competences that the test takers are expected to be able to handle?	These are clearly set out in the British Council/EAQUALS Core Inventory.
8. After reading the scale for Strategic Competence in Table A3, indicate and justify at which level(s) of the scale the examination should be situated.	**Level A1-C** **Justification (incl. reference to documentation)** • The items, themes and foci of the input texts were drawn from the CEFR, while the areas of linguistic competence are based on a Core Inventory which itself is very much driven by the CEFR. • Quality control of items ensures that the item and task writers continue to meet the expectations of the specifications. • The specifications are created in such a way as to encourage interaction between the item writers and the quality assurance team. • Standard setting acts to triangulate this evidence

Form A19: Aspects of Language Competence in Reception (continued)

A4.2 Interaction

Those CEFR scales most relevant to Interaction have been used to create Table A4 which can be referred to in this section. Table A4 does not include any descriptors for "plus levels". The original scales consulted, some of which do define plus levels, include:

Linguistic Competence

- General Linguistic Range English: page 110
- Vocabulary Range English: page 112
- Vocabulary Control English: page 112
- Grammatical Accuracy English: page 114

Socio-linguistic Competence

- Socio-linguistic Appropriateness English: page 122

Pragmatic Competence

- Flexibility English: page 124
- Turntaking English: page 124
- Spoken Fluency English: page 129
- Propositional Precision English: page 129

Strategic Competence

- Turntaking (repeated) English: page 86
- Cooperating English: page 86
- Asking for Clarification English: page 87
- Compensating English: page 64
- Monitoring and Repair English: page 65

Linguistic Competence	Short description and/or reference
1. What is the range of lexical and grammatical competence that the test takers are expected to be able to handle?	This is clearly set out in the British Council/EAQUALS Core Inventory.
2. What is the range of phonological and orthographic competence that the test takers are expected to be able to handle?	This is clearly set out in the British Council/EAQUALS Core Inventory.
3. After reading the scales for Range and Accuracy in Table A4, indicate and justify at which level(s) of the scale the examination should be situated.	**Levels A1-C** **Justification (incl. reference to documentation)** • The items, themes and foci of the input texts were drawn from the CEFR, while the areas of linguistic competence are based on a Core Inventory which itself is very much driven by the CEFR. • Quality control of items ensures that the item and task writers continue to meet the expectations of the specifications. • The specifications are created in such a way as to encourage interaction between the item writers and the quality assurance team. • Standard setting acts to triangulate this evidence • Standard setting acts to triangulate this evidence • See Table A4
Socio-linguistic Competence	
4. What are the socio-linguistic competences that the test takers are expected to be able to handle: linguistic markers, politeness conventions, register, adequacy, dialect/accent, etc.?	This is clearly set out in the British Council/EAQUALS Core Inventory.
5. After reading the scale for Socio-linguistic Competence in Table A4, indicate and justify at which level(s) of the scale the examination should be situated.	**Levels A1-C** • The items, themes and foci of the input texts were drawn from the CEFR, while the areas of sociolinguistic competence are based on a Core Inventory which itself is very much driven by the CEFR. • Quality control of items ensures that the item and task writers continue to meet the expectations of the specifications. • The specifications are created in such a way as to encourage interaction between the item writers and the quality assurance team. • Standard setting acts to triangulate this evidence • Standard setting acts to triangulate this evidence • See Table A4

Pragmatic Competence	Short description and/or reference
6. What are the pragmatic competences that the test takers are expected to be able to handle: discourse competences, functional competences?	This is clearly set out in the British Council/EAQUALS Core Inventory.
7. After reading the scale for Fluency in Table A4, indicate and justify at which level(s) of the scale the examination should be situated.	**Levels A1-C** **Justification (incl. reference to documentation)** • The items, themes and foci of the input texts were drawn from the CEFR, while the areas of pragmatic competence are based on a Core Inventory which itself is very much driven by the CEFR. • Quality control of items ensures that the item and task writers continue to meet the expectations of the specifications. • The specifications are created in such a way as to encourage interaction between the item writers and the quality assurance team. • Standard setting acts to triangulate this evidence • Standard setting acts to triangulate this evidence • See Table A4

Form A20: Aspects of Language Competence in Interaction (part)

Strategic Competence	Short description and/or reference
8. What are the interaction strategies that the test takers are expected to be able to handle? The discussion in CEFR 4.4.3.5 might be of help as a reference.	This is clearly set out in the British Council/EAQUALS Core Inventory.
9. After reading the scale for Interaction in Table A4, indicate and justify at which level(s) of the scale the examination should be situated.	**Levels A1-C** **Justification (incl. reference to documentation)** • The items, themes and foci of the input texts were drawn from the CEFR, while the areas of strategic competence are based on a Core Inventory which itself is very much driven by the CEFR. • Quality control of items ensures that the item and task writers continue to meet the expectations of the specifications. • The specifications are created in such a way as to encourage interaction between the item writers and the quality assurance team. • Standard setting acts to triangulate this evidence • Standard setting acts to triangulate this evidence • See Table A4

Form A20: Aspects of Language Competence in Interaction (continued)

A4.3 Production

Those CEFR scales most relevant to Production have been used to create Table A5, which can be referred to in this section. Table A5 does not include any descriptors for "plus levels". The original scales consulted, some of which do define plus levels, include:

Linguistic Competence

- General Linguistic Range English: page 110
- Vocabulary Range English: page 112
- Vocabulary Control English: page 112
- Grammatical Accuracy English: page 114

Socio-linguistic Competence

- Socio-linguistic Appropriateness English: page 122

Pragmatic Competence

- Flexibility English: page 124
- Thematic Development English: page 125
- Cohesion and Coherence English: page 125
- Spoken Fluency English: page 129
- Propositional Precision English: page 129

Strategic Competence

- Planning English: page 64
- Compensating English: page 64
- Monitoring and Repair English: page 65

Linguistic Competence	Short description and/or reference
1. What is the range of lexical and grammatical competence that the test takers are expected to be able to handle?	This is clearly set out in the British Council/EAQUALS Core Inventory.
2. What is the range of phonological and orthographic competence that the test takers are expected to be able to handle?	This is clearly set out in the British Council/EAQUALS Core Inventory.
3. After reading the scales for Range and Accuracy in Table A5 indicate and justify at which level(s) of the scale the examination should be situated.	**Levels A1-C** **Justification (incl. reference to documentation)** • The items, themes and foci of the input texts were drawn from the CEFR, while the areas of strategic competence are based on a Core Inventory which itself is very much driven by the CEFR. • Quality control of items ensures that the item and task writers continue to meet the expectations of the specifications. • The specifications are created in such a way as to encourage interaction between the item writers and the quality assurance team. • Standard setting acts to triangulate this evidence • Standard setting acts to triangulate this evidence • See Table A5
Socio-linguistic Competence	**Short description and/or reference**
4. What are the socio-linguistic competences that the test takers are expected to be able to handle: linguistic markers, politeness conventions, register, adequacy, dialect/accent, etc.?	This is clearly set out in the British Council/EAQUALS Core Inventory.
5. After reading the scale for Socio-linguistic Competence in Table A5, indicate and justify at which level(s) of the scale the examination should be situated.	**Levels A1-C** **Justification (incl. reference to documentation)** • The items, themes and foci of the input texts were drawn from the CEFR, while the areas of sociolinguistic competence are based on a Core Inventory which itself is very much driven by the CEFR. • Quality control of items ensures that the item and task writers continue to meet the expectations of the specifications. • The specifications are created in such a way as to encourage interaction between the item writers and the quality assurance team. • Standard setting acts to triangulate this evidence • Standard setting acts to triangulate this evidence • See Table A5

Form A21: Aspects of Language Competence in Production (part)

Pragmatic Competence	Short description and/or reference
6. What are the pragmatic competences that the test takers are expected to be able to handle: discourse competences, functional competences? The lists in CEFR 5.2.3 might be of help as a reference.	This is clearly set out in the British Council/EAQUALS Core Inventory.
7. After reading the scale for Pragmatic Competence in Table A5, indicate and justify at which level(s) of the scale the examination should be situated.	**Levels A1-C** **Justification (incl. reference to documentation)** • The items, themes and foci of the input texts were drawn from the CEFR, while the areas of strategic competence are based on a Core Inventory which itself is very much driven by the CEFR. • Quality control of items ensures that the item and task writers continue to meet the expectations of the specifications. • The specifications are created in such a way as to encourage interaction between the item writers and the quality assurance team. • Standard setting acts to triangulate this evidence • Standard setting acts to triangulate this evidence • See Table A5
Strategic Competence	**Short description and/or reference**
8. What are the production strategies that the test takers are expected to be able to handle? The discussion in CEFR 4.4.1.3 might be of help as a reference.	This is clearly set out in the British Council/EAQUALS Core Inventory.
9. After reading the scale for Strategic Competence in Table A5, indicate and justify at which level(s) of the scale the examination should be situated.	**Levels A1-C** **Justification (incl. reference to documentation)** • The items, themes and foci of the input texts were drawn from the CEFR, while the areas of sociolinguistic competence are based on a Core Inventory which itself is very much driven by the CEFR. • Quality control of items ensures that the item and task writers continue to meet the expectations of the specifications. • The specifications are created in such a way as to encourage interaction between the item writers and the quality assurance team. • Standard setting acts to triangulate this evidence • Standard setting acts to triangulate this evidence • See Table A5

Form A21: Aspects of Language Competence in Production (continued)

TABLE A4: RELEVANT QUALITATIVE FACTORS FOR SPOKEN INTERACTION

	LINGUISTIC RANGE: Edited from General Linguistic Range; Vocabulary Range, Flexibility	**LINGUISTIC ACCURACY: Edited from Grammatical Accuracy and Vocabulary Control**	**SOCIO-LINGUISTIC: Edited from Socio-linguistic Appropriateness**	**FLUENCY: Fluency, Flexibility**	**INTERACTION: Edited from Turntaking, Cooperating, Asking for Clarification**
C2	*Shows great flexibility reformulating ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis, to differentiate and to eliminate ambiguity. Also has a good command of idiomatic expressions and colloquialisms.*	*Maintains consistent grammatical control of complex language, even while attention is otherwise engaged (e.g. in forward planning, in monitoring others' reactions).*	*Appreciates fully the socio-linguistic and sociocultural implications of language used by speakers and can react accordingly.* *Can mediate effectively between speakers of the target language and that of his/her community of origin taking account of sociocultural and socio-linguistic differences.*	*Can express him/herself spontaneously at length with a natural colloquial flow, avoiding or backtracking around any difficulty so smoothly that the interlocutor is hardly aware of it.*	*Can interact with ease and skill, picking up and using non-verbal and intonational cues apparently effortlessly. Can interweave his/her contribution into the joint discourse with fully natural turntaking, referencing, allusion making etc.*
C1	*Has a good command of a broad range of language allowing him/her to select a formulation to express him/herself clearly in an appropriate style on a wide range of general, academic, professional or leisure topics without having to restrict what he/she wants to say.*	*Consistently maintains a high degree of grammatical accuracy; errors are rare, difficult to spot and generally corrected when they do occur.*	*Can use language flexibly and effectively for social purposes, including emotional, allusive and joking usage.*	*Can express him/herself fluently and spontaneously, almost effortlessly. Only a con-ceptually difficult subject can hinder a natural, smooth flow of language.*	*Can select a suitable phrase from a readily available range of discourse functions to preface his remarks in order to get or to keep the floor and to relate his/her own contributions skil-fully to those of other speakers.*
B2	*Has a sufficient range of language to be able to give clear descriptions, express viewpoints on most general topics, without much conspicuous searching for words, using some complex sentence forms to do so.*	*Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstanding, and can correct most of his/her mistakes.*	*Can with some effort keep up with and contribute to group discussions even when speech is fast and colloquial.* *Can sustain relationships with native speakers without unintentionally amusing or irritating them or requiring them to behave other than they would with a native speaker.*	*Can adjust to the changes of direction, style and emphasis normally found in conversation.* *Can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he or she searches for patterns and expressions, there are few noticeably long pauses.*	*Can initiate discourse, take his/her turn when appropriate and end conversation when he/she needs to, though he/she may not always do this elegantly. Can help the discussion along on familiar ground confirming comprehension, inviting others in, etc.*
B1	*Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocutions on topics such as family, hobbies and interests, work, travel, and current events.*	*Uses reasonably accu¬rately a repertoire of frequently used "routines" and patterns asso¬ciated with more predictable situations.*	*Can perform and respond to basic language functions, such as information exchange and requests and express opinions and attitudes in a simple way. Is aware of the salient politeness conventions and acts appropriately.*	*Can exploit a wide range of simple language flexibly to express much of what he/she wants.* *Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production.*	*Can initiate, maintain and close simple face-to-face conversation on topics that are familiar or of personal interest. Can repeat back part of what someone has said to confirm mutual understanding.*

Form A21: Aspects of Language Competence in Production (continued)

TABLE A4: RELEVANT QUALITATIVE FACTORS FOR SPOKEN INTERACTION

	LINGUISTIC RANGE: Edited from General Linguistic Range; Vocabulary Range, Flexibility	LINGUISTIC ACCURACY: Edited from Grammatical Accuracy and Vocabulary Control	SOCIO-LINGUISTIC: Edited from Socio-linguistic Appropriateness	FLUENCY: Fluency, Flexibility	INTERACTION: Edited from Turntaking, Cooperating, Asking for Clarification
A2	*Uses basic sentence patterns with memorised phrases, groups of a few words and formulae in order to communicate limited information in simple everyday situations.*	*Uses some simple structures correctly, but still systemati¬cally makes basic mistakes.*	*Can handle very short social exchanges, using everyday polite forms of greeting and address. Can make and respond to invitations, apologies etc.*	*Can make him/herself understood in very short utterances, even though pauses, false starts and reformulation are very evident. Can expand learned phrases through simple recombinations of their elements.*	*Can indicate when he/she is following but is rarely able to understand enough to keep conversation going of his/her own accord.* *Can ask for attention.*
A1	*Has a very basic repertoire of words and simple phrases related to personal details and particular concrete situations.*	*Shows only limited grammatical control of a few simple grammatical structures and sentence patterns in a memorised repertoire.*	*Can establish basic social contact by using the simplest everyday polite forms of: greetings and farewells; introductions; saying please, thank you, sorry etc.*	*Can manage very short, isolated, mainly pre-packaged utterances, with much pausing to search for expressions, to articulate less familiar words, and to repair communication.*	*Can interact in a simple way but communication is totally dependent on repetition, rephrasing and repair.*

TABLE A5: RELEVANT QUALITATIVE FACTORS FOR PRODUCTION

	LINGUISTIC RANGE: General Linguistic Range; Vocabulary Range	LINGUISTIC ACCURACY: Grammatical Accuracy, Vocabulary Control, Phonological Control	SOCIO-LINGUISTIC: Socio-linguistic Appropriateness	PRAGMATIC: Fluency, Flexibility	PRAGMATIC: Thematic Development, Propositional Precision, Coherence and Cohesion	STRATEGIC: Compensating, Monitoring and Repair
C2	*Shows great flexibility reformulating ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis, to differentiate and to eliminate ambiguity. Also has a good command of idiomatic expressions and colloquialisms.*	*Maintains consistent grammatical control of complex language, even while attention is otherwise engaged (e.g. in forward planning, in monitoring others' reactions).*	*Appreciates fully the socio-linguistic and sociocultural implications of language used by speakers and can react accordingly.*	*Can express him/herself spontaneously at length with a natural colloquial flow, avoiding or backtracking around any difficulty so smoothly that the interlocutor is hardly aware of it.*	*Can create coherent and cohesive discourse making full and appropriate use of a variety of organisational patterns and a wide range of connectors and other cohe¬sive devices.*	*Can substitute an equivalent term for a word he/she can't recall so smoothly that it is scarcely noticeable.*
C1	*Has a good command of a broad range of language allow¬ing him/her to select a formulation to express him/herself clearly in an appropriate style on a wide range of general, academic, professional or leisure topics without having to restrict what he/she wants to say.*	*Consistently maintains a high degree of grammatical accuracy; errors are rare, difficult to spot and generally corrected when they do occur.*	*Can use language flexibly and effectively for social purposes, including emotional, allusive and joking usage.*	*Can express him/herself fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language.*	*Can produce clear, smoothly flowing, well-structured speech, showing controlled use of organisational patterns, connectors and cohesive devices.* *Can give elaborate descriptions and narratives, integrating sub themes, developing particular points and rounding off with an appropriate conclusion.*	*Can backtrack when he/she encounters a difficulty and reformulate what he/she wants to say without fully interrupting the flow of speech.*

TABLE A5: RELEVANT QUALITATIVE FACTORS FOR PRODUCTION

	LINGUISTIC RANGE: General Linguistic Range; Vocabulary Range	LINGUISTIC ACCURACY: Grammatical Accuracy, Vocabulary Control, Phonological Control	SOCIO-LINGUISTIC: Socio-linguistic Appropriateness	PRAGMATIC: Fluency, Flexibility	PRAGMATIC: Thematic Development, Propositional Precision, Coherence and Cohesion	STRATEGIC: Compensating, Monitoring and Repair
B2	*Has a sufficient range of language to be able to give clear descriptions, express viewpoints on most general topics, without much conspicuous searching for words, using some complex sentence forms to do so.*	*Shows a relatively high degree of grammatical control. Does not make errors which cause misun¬derstanding, and can correct most of his/her mistakes.*	*Can express him or herself appropriately in situations and avoid crass errors of formulation.*	*Can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he or she searches for patterns and expressions, there are few noticeably long pauses.*	*Can develop a clear description or narrative, expanding and supporting his/her main points with relevant supporting detail and examples.* *Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse, though there may be some "jumpiness" in a long contribution.*	*Can use circumlocution and paraphrase to cover gaps in vocabulary and structure.* *Can make a note of "favourite mistakes" and consciously monitor speech for it/them.*
B1	*Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocutions on topics such as family, hobbies and interests, work, travel, and current events.*	*Uses reasonably accurately a repertoire of frequently used "routines" and patterns associated with more predictable situations.*	*No descriptor available*	*Can exploit a wide range of simple language flexibly to express much of what he/she wants. Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production.*	*Can link a series of shorter, discrete simple elements in order to reasonably fluently relate a straightforward narrative or description as a linear sequence of points.*	*Can use a simple word meaning something similar to the concept he/she wants to convey and invites "correction".* *Can start again using a different tactic when communication breaks down.*
A2	*Uses basic sentence patterns with memorised phrases, groups of a few words and formulae in order to communicate limited information in simple everyday situations.*	*Uses some simple structures correctly, but still systematically makes basic mistakes.*	*No descriptor available*	*Can make him/herself understood in very short utterances, even though pauses, false starts and reformulation are very evident. Can expand learned phrases through simple recombinations of their elements.*	*Can link groups of words with simple connectors like "and", "but" and "because".*	*No descriptor available*
A1	*Has a very basic repertoire of words and simple phrases related to personal details and particular concrete situations.*	*Shows only limited control of a few simple grammatical structures and sentence patterns in a memorised repertoire.*	*No descriptor available*	*Can manage very short, isolated, mainly pre-packaged utterances, with much pausing to search for expressions, to articulate less familiar words, and to repair communication.*	*Can link words or groups of words with very basic linear connectors like "and" or "then".*	*No descriptor available*

Section A5: Specification: Outcome of the Analysis (Chapter 4)

Form A23 provides a graphic profile of the coverage of the examination in relation to CEFR categories and levels. It should be completed at the end of the Specification process.

C								
B2								
B1								
A2								
A1								
	Knowledge	Listening	Reading	Writing	Speaking	Sociolinguistic	Pragmatic	Linguistic

Form A23: Graphic Profile of the Relationship of the Examination to CEFR Levels

Initial estimation of overall CEFR level
Short rationale, reference to documentation. If this form presents a different conclusion to the initial estimation in Form A8, please comment on the principal reasons for the revised view. The evidence presented here and in the rest of the report indicates that the test access is language across all levels. The specifications for the test were created with the CEFR as its basis. The specifications are constantly reviewed and reflected on by the quality assurance team and the item writers. In addition, all items and tasks are extensively trialled before usage in the test. The standard-setting section of this report shows that there is a clear link between the various boundary points and the CEFR, as claimed. Finally, the validation section of the report offers evidence that the test is robust accurate and reliable. Is evident also supports and justifies playing as the test is likely to function in a consistent way.

APPENDIX 2: APTIS WRITING PAPER SCALES

Task 2 Scale

5	Likely to be above the A2 level
4 [A2.2]	Clearly defined sentences all on topic. Mostly accurate grammar with few serious errors of vocabulary usage (i.e. appropriateness and spelling). The text organisation is completely appropriate for task. Attempts at textual cohesion and accurate punctuation.
3 [A2.1]	There are some serious issues with grammar and vocabulary usage. However, the meaning still clear. Text written in complete sentences, organised appropriately for the text form and mostly accurate punctuation.
2 [A1.2]	Numerous serious errors of grammar and vocabulary usage which make the text sometimes difficult to follow. A series of phrases, not sentences. Poor punctuation.
1 [A1.1]	There is too little language or the usage is so poor that the text is almost impossible to follow. There is no clear structure.
0	Little or no meaningful language or the work is off-topic.

Task 3 Scale

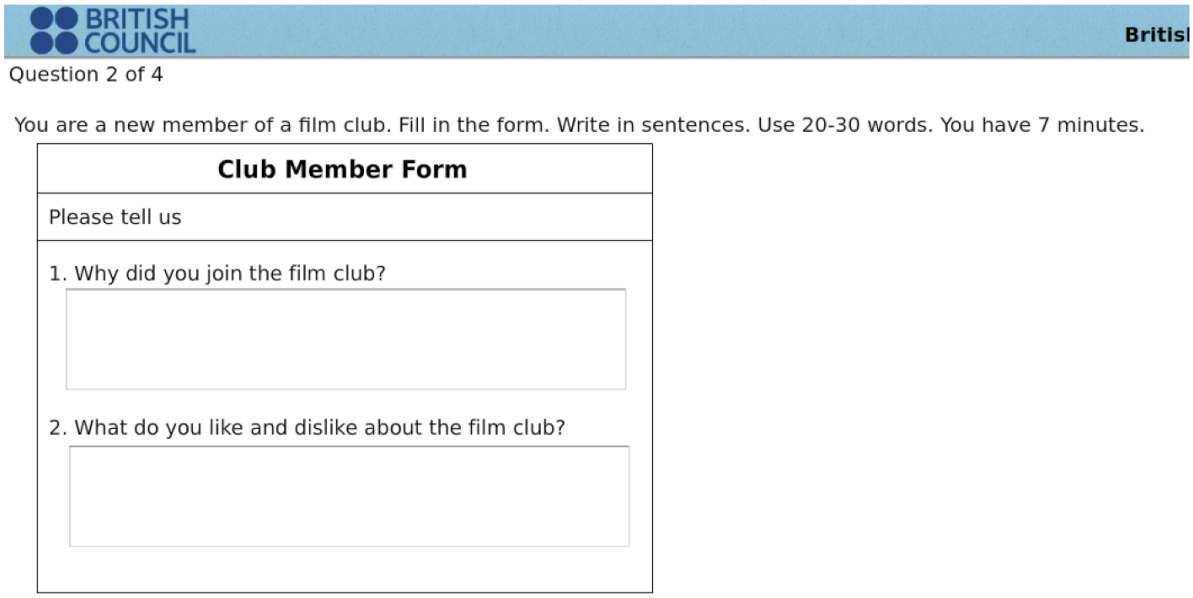
5	Likely to be above the B1 level
4 [B1.2]	Replies fully to each piece of input The grammar is appropriate to B1 and is mostly accurate, while there is a good range of vocabulary on general topics. Some errors but these don't impede communication Cohesive and coherent text appropriately using an appropriate range of linguistic devices. Few if any punctuation or spelling errors.
3 [B1.1]	Replies well to at least two of the input texts. An adequate range of grammar used with no major errors which impact on understanding. There is good control of elementary vocabulary, though evidence of some major errors when expressing unfamiliar or complex topics Cohesive and coherent text adequately using a range of linguistic devices. Spelling and/or punctuation errors do not impede communication.
2 [A2.2]	Replies to at least two of the input texts. Many errors which make the text sometimes difficult to follow. Narrow lexical repertoire, here again, frequent errors make the message difficult to follow. Some effort to use connecting devices though not always consistent. Errors, including punctuation and spelling, make the text difficult to follow.
1 [A2.1]	Does not reply to more than one input. There is little language with such poor control as to make the text almost impossible to follow without considerable effort. Very basic for everyday vocabulary. Lacks cohesion and/or uses linguistic devices inappropriately. Spelling and punctuation errors make the text almost impossible to follow.
0	Little or no meaningful language or the work is off-topic.

Task 4 Scale

5	Likely to be above the B2 level
4 [B2.2]	Task fulfilled in terms of appropriateness of register (i.e. two distinct registers used in the different messages written). Evidence of a clear, assured and precise use of a broad range of grammatical forms used. A good command of a broad lexicon. Good use of idiomatic expressions and no impeding errors of grammar or lexis. Few if any errors of cohesion or coherence.
3 [B2.1]	Task partially fulfilled in terms of appropriateness of register (i.e. fully appropriate register used in one of the two different messages written). An adequate range of grammatical forms used, with no impeding errors. A good range of lexis with a high level of accuracy. Errors do not affect the message. Cohesive and coherent text adequately using a range of linguistic devices. Spelling and/or punctuation errors evident but these do not affect the message.
2 [B1.2]	Task not fulfilled in terms of appropriateness of register (i.e. appropriate register not used in either of the two different messages written). A relatively narrow range of grammatical forms used, with some impeding errors. The lexical range adequate for the description of situations relating to him/herself. Some errors which tend to make understanding difficult. Attempts to use linguistic devices though not always consistent. Errors, including punctuation and spelling, can make understanding difficult.
1 [B1.1]	Task not fulfilled in terms of appropriateness of register (i.e. no evidence of awareness of register). A limited range of grammatical forms and vocabulary used and not always with sufficient accuracy. Errors may make the text difficult to follow Lacks systematic cohesion and/or uses linguistic devices inappropriately. Spelling and punctuation errors can make understanding difficult.
0	Clearly below the B level; work is off topic

APPENDIX 3: TASKS AND SCRIPTS INCLUDED IN THE WRITING EVENT

Task 2 [A2]



Student ID 93680513

I joined the club because it sounded interesting and it was exactly what I was looking for!

How can anyone dislike something about this club? It's almost impossible! The things I like more are the way we are learning and the activities we do! The only thing I dislike it's some of the tecnical problems!

Student ID 93680511

I joined the club because there's activities i'm interested in participate and a lot of my friends recommended this club to me.

I like the friendship that we can create by joining a club like this, what i don't like are the little problems that sometimes i have to carry on because of some things that i've done in the club.

Student ID 93680516

I like all kind of films but specially drama ones. "Slumdog Millionaire" is my favourite one.

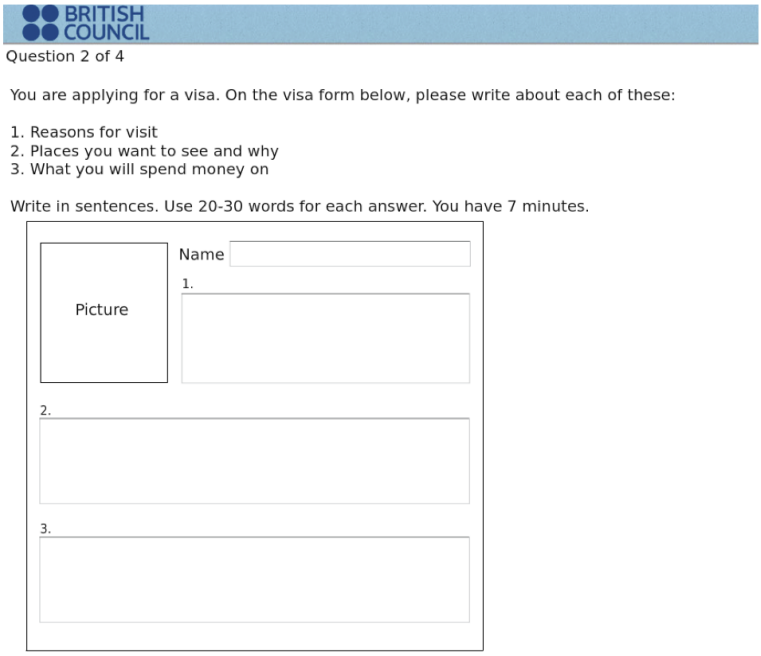
I'm sorry but already have something to do after the Saturday's film showing...

Student ID 93680572

I join the club to improve My English to help me to emigrate to Austerlia. i want to discover more Land and area.

I like the club because it helps me to improve my Enhlish and teah me more in English. Idislike it because it sometimes comes boring.

XXXXX



Student ID 93683074

I am going in businus trip. Have meeting with Factory pump to use it in the project i am working on it.

It is my first time to me to visit London so of course i want to see roal palce alos to visit the all the musiums in london. also i heared that the will make a big conference for the new technolgy avilable in the pumps cntrol may i can go there.

I really want buy as much as i can to buy a souvners and gifts from London.also i will try to catch if there are good shows of threater are there, i will try to go. may be if i have some time i will try to go to another around in Englind

Student ID 93683062

One week trip to UK for vacation.

Planning to visit London, Lake Area, and Scotland, because these are crowned as most recommended place to go for people%u2019s first trip to UK.

I will most spend money on accommodation, foods, and transports.

Student ID 93683094

I would like to visit, becouse i want to see different culture, meet new people and I love adventures.

I want to see every places, which are recommended in internet. I want to visit bigest city, to see history places.

I want to spend my money for transport, food, suvenirs.

Student ID 93683092

I want to go Mexico with my friends in the Summer because there are very beautiful island and beach.Second reason is I want to meet my best friend ,because She Studies there.

First time I am going to my friend because the hotel is very expensive and my friend only lives there. I am going to visit ancient captal and see famous building,historic town, museums, and go swimming and shopping.In the end I am going to visit are gallary because I like painting pictures very mush. I am also going to have Mexico festival with my best friend.

I am not sure,but I will spend money on buying clothes, playing on the beach,having party,and I will buy presents about my parents. Maybe I think,I shound eating in a restaurant.But I am going to save money.

Task 2 [A2]

Here is a postcard of my town.
Please send me a postcard from
your town. What size is your town?
What is the nicest part of your town?
Where do you go in the evenings?

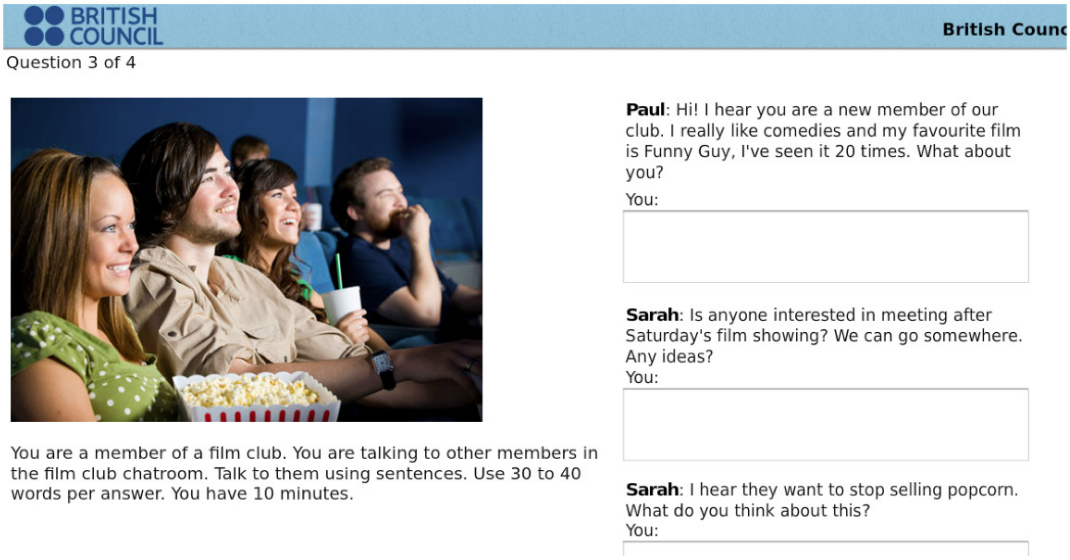
Sam

Write Sam a postcard. Answer the questions. Write 25-35 words.

Dear Sam,
I lived in a small town, although it was small but lovely.
People lived in my town are friendly and nice, they always help each other.
I think that's the nicest part of my town. I hope you can come here.
By the way I'm not went out in evenings.

Love

Task 3 [B1]



Student ID 93680513

Hello Paul! Yes, I'm. Wow me too! I really enjoy comedies. But my favorite film isn't a comedy one. It's a horror one, called Paranormal Activity. Have you ever seen it? It's really good and scary! I bought the DVD so I could watch it anytime!

Yes I'm! We could go to a shopping mall and see the new clothes that just arrived! But I think the boys woudn't like that! Or we could make a long walk at the beach and see the ocean! And just relax a little.

What?! I didn't heard that! How can they do that? When I go to the cinema I really enjoy my bucket of hot popcorns! I think they aren't going to do that, because that way they will loose money!

Student ID 93680511

Hey! I joined the club recently has you know. My favourite genre of films are the horror ones and my favourite film is Friday the 13th. I watched it for the first time when i was a little kid and since there i love it!

What a great idea! There's an old film store in front of the cinema that sells really good films for a great price. We can go there and check if there's something of our interest. I swear you won't be disappointed.

That's just stupid! Who doesn't love to watch a good film accompained by some really tasty bucket of popcorn? If they stop selling them, people are you going to stay at home and watch a film because they can eat whatever they want to.

Student ID 93680516

It would really be a bad idea, i think. I go to the cinema to watch the movies but i must confess i also go there for the delicious popcorns.

I'm sorry but already have something to do after the Saturday's film showing...

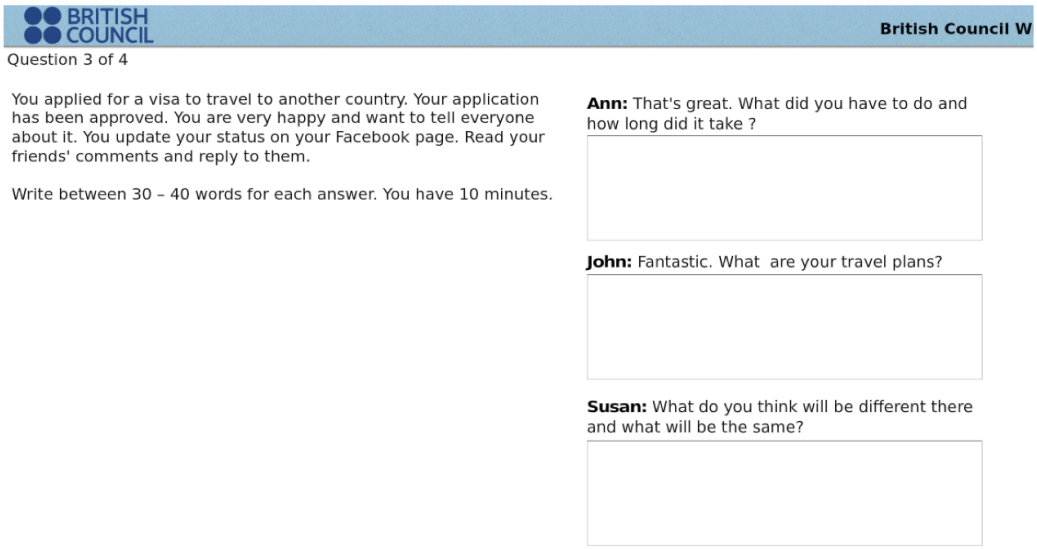
It would really be a bad idea, i think. I go to the cinema to watch the movies but i must confess i also go there for the delicious popcorns.

Student ID 93680572

Really because i like the comedy movies also, and my favorete Film is The Mask, i watched more than 20 times also My favorte actress is Jim Cary

I think we can go out for a restuart or having a walk around , or may be a cafeshop . OHH.. i Know a cafe name hipark i think it is good.they sell a very good popcorn.

i think it will be very boring.



Student ID 93683074

Really I am going in a business trip , you know just for work. but i think i may stay for more time to go around Englind to see and visit many places as much as i can. My business trip will take 1 week but iwill stay for another week.

Sure i am going to see the roal palce and to move around in London as much as i can ,if i have some free time i will try to move around all england as much as i can.

for sure the roal family will be the same, but but what will be different the new molls that are built in south of london. this molls is huge and large in this area. i believe i will make to much shopping.

Student ID 93683062

Just filled in a form, and submit relevant documents they required. It took me around 2 weeks to get the visa upon the date I submitted the documents.

The first place to go is, of course, London. I am going to spend 3 days there, and then travel to Lake Area, and then Scotland, and then back to London again.

I heard the landscape in Lake Area and Scotland is very much different from it is here. This is the main reason triggered my trip to UK. As for London, I guess the thing most different from here would be the lifestyle. We will see'

Student ID 93683094
I just apply for visa and give them the document they want from me. The procedure for visa takes 1 month. I have no plans. Just take the airplane and jump into adventures I don't know. I will tell you when I come back. I think everything is different there.
Student ID 93683092
I had to apply my visa and buy plane tickets .But I found buying my cheap ticket on the Intenet for long time because it was very expensive. Other friend bought their tickes for long time too.I took It for 2 weeks My travel plans are meeting my best friend,going swimming and shopping and sightseeing ,visiting art gallary.In the end I am going to go wine shops ,because my dad likes red wine very much. I think ,it is very important about visiting other one country. I think ,their language is a littel different from English.Maybe there life, having foods,geting up, going to bed are different with our county life.I am not sure ,Althogh they are Americans ,there psychology likes our country.

Additional Task [B1]

- Your English Teacher has asked you to write a story.
- Your story must have the following title:

The most important day of my life

The most important day of my life.
During a lifetime there is so many days you could call, the most important day of your life". It could be the day you chose wich school you are going to, or what you want to work with the rest of your life. Another important day is when you get married, or you chose where to live.
But most of all it must be a very important day when you give birth to a child. I think that changes everything you have been doing until then. Than you have to realice that somebody are more important than yourself.

Task 4 [B2]

Student ID 93680513

Hello Mary! Have you heard the news? The main hall of our film club will be closed for painting and we have to see the films on DVD in the lounge! The maximum of seats will be of 25 per showing! I don't think this is right, because they are a lot of people in our club that want to watch the films! I'm really upset about this! I think they should rent a bigger room and that way all of us would watch the films!Tell me what you think about this! XOXO Gabriela

Good afternoon, I'm a member of your film club. I heard that the main hall of the film club will be closed due to painting. And the members will have to see the films on DVD in the lounge, with the maximum of seats being 25 per showing. I think we should think in other solutions, like for example rent a bigger room and we all could fit there. Because a lot of members are upset (including me) and don't want to watch the films there because we simply don't have any space. We understand that it needed to be painted but we can always suggest other possible alternatives. Sincerely, Gabriela.

Student ID 93680511

Hey John, do you heard the news about the film club? They're going to close the main hall for painting, that's not right! We need to get a place to watch our films and it has to be really big. I remembered once you've said that you had a house here in the city that was completely empty. What do you think if we start doing our movie-marathons there?

Dear Film Club Manager Every single members of the film club have heard about the terrible news and we are shocked. Don't you think we should get a better solution instead of showing the films in the lounge? It's a really small place and the club has like 150 members or even more. Besides, we need a bigger screen than the one that's in the lounge. I talked to a friend of mine and he is as much indignated as me. He has a house completely empty house and he agreed to borrow the house until the main hall is finished. What do you think about the idea? We can take the chairs and the screen from our film room and put them in there. I think it's perfect. Regards, Anthony

Student ID 93680516

Hi Sara! When I saw in the noticed that a maximum of 25 seats will be available for watch the films I was really sad because, as I always get late, i don't think that, when I come to see the movie, any seats will be available for me. It's too bad but I don't think i'm going there anymore.

Manager of the club, i don't think it was a really good idea to close the hall for painting, speacially in this time of the year. I think you already know that, as the winter is coming, the hall would be warmer than the lounge and i don't believe people will be uncomfortable just to see a film (doesn't matter how good it might be). I work until the hour the films start so, even if i go really quickly over there, i always get a little late and, with only 25 seats available, i don't think i can have mine. In my opinion, you should replace the hall for another place but the lounge. Somewhere warmer and, if possible, bigger. This way people would feel more comfortable and, with more seats available, i would have more chances to find one for me when i get there. Thank you, Adriana.

Student ID 93680572

Dear Michael : How are You? how are things going? i have a bad news for you.the main hall is under renewing . and they are going to change to into DVD in the Club with max. 25 chair this means that not all of us can meet together. ithink we have to meet togther in cenima Metro every sunday night so we can see each other until the finis the big hall. or even let's see if any of or friends have any other ideas.

Dear Sir: I am writing this mail just to tell how sorry i am because of closing the big hall i fell .me and alot of the members who are meeting every weak in this hall and share alot of Fun and nice time there. it was a very unhappy thing for closing that hall. I believe i have avery good idea for you.what if try to split the big hall in two area and you can work on one , and when you finish move to other one. so always have an areato meet together.

BRITISH COUNCIL British Coun

Question 4 of 4

You applied for a visa to visit another country. You received your visa but this time there were problems getting it.

Part 1

Post an update on your Facebook page telling your friends the problems. Write about 50 words. You have 10 minutes.

Part 2

You are upset about the problems with the visa process. You decide to write an email to the embassy to complain and suggesting they improve their service. Write about 100 words. You have 20 minutes.

Student ID 93683074

I have a problem with my visa. My name was simillar too much simillar to guy who is criminal. they though that i am that guy becuse of these.you know should try to check the photos. they will know that i am not him.

Dear sir : I woulld to tell you about my problem. the embassy reject my visa because my name was similar to guy who made a lot of crimes. i think if you try to check the photos you dicover that i am not that man.i think if you use computer program that help to identiy the photos, it will help to much to check on the people, also if can make some investigation through our police department,they will help you more and give you more data about me.you. i am good man who earn his money from his work and i have never made any crimes.you should have check more pleas nect time . even with other people. your sincerly Moataz Mohamed

Student ID 93683062

Can you imaging that the document I was told to prepare was not the correct one I was supposed to submit? I couldn%u2019t understand what happened, but the thing I do know is there will not be enough of time for me to get my visa before I leave for UK, I couldn't make the trip I have planned for months.

To Whom It May Concern: I was informed yesterday morning by Sisley Zhu of the failure in processing my visa to UK, and the reason was that the annual income supporting was not the correct document I should submit. I couldn't understand this because the document check list I get from your office weeks ago indicates very clearly that the annual income supporting is one of the correct documents. I therefore came to your office this morning with this document check list, however, I didn't get any reasonable answers from you, except for being asked to wait for the receptionist for almost one hour. Can you at least ensure that the information you deliver to the customers is accurate and consistent? I would also grateful if you could get back to me to give me a reasonable answer. Best regards, Jiang Lin

Student ID 93683094

my friends...I'm so sad. there are problems with my visa. The people from the embassy think that I don't have the money for my travell and want give them a bank statement to prove my finance status.

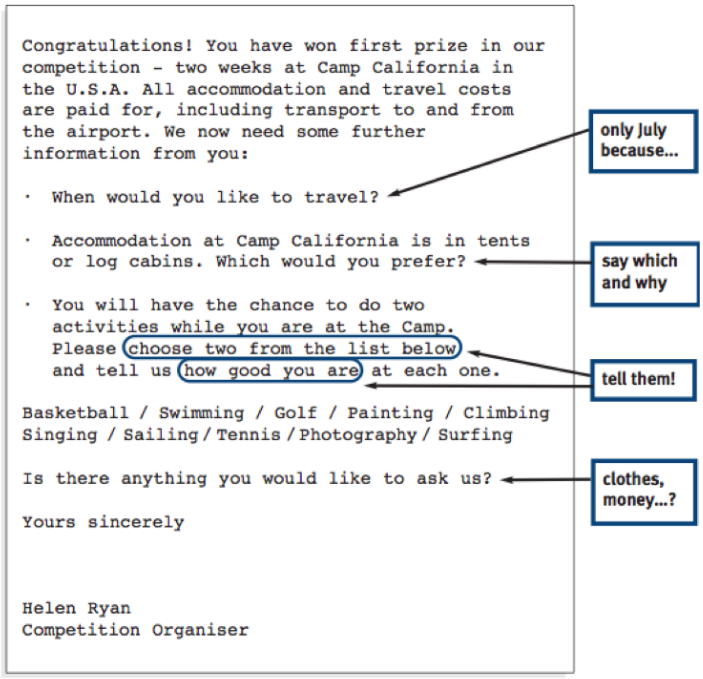
Dear Embassador, I would like to make some complain and suggestions about services of the embassy. It took you so loog to take a decision for my applly - I think you need to improve that.

Student ID 93683092

I have problems about applying visa because I want to stay there for 2 weeks but there only 7days.I think this dates are very short about travelling .Unluckily this year a lot of people offer visa. What can I do?

Hello Mr (Mrs) Be I have problems with travelling dates.Before you said what I could travel for long times allow my deciding.I can't understand why your saying is different with before.I think ,your moodying is the worst I've been offer my visa.Nowl got very angry with my husband.If you can help me ,I will thanks to you.I am really sorry about my bad complaining.If you allow my planning, I am sure maybe I will have very nice and exciting travelling. Take care of yourself. 13 11 2011

Additional Task [B2]



Example answer

Competition Organiser
Helen Ryan

Thank you very much for the letter that telling me I won first prize in the competition. I am so glad and I am going to write some information that you need from me. First of all, I would like to travel only July because It is due to my job. And about accommodation I would prefer log cabins to tents. I have never stayed log cabins so it would be good chance to me. In your letter, you mentioned that I have the chance to do two activities. I would choose Sailing and Photography. However, I am absolutely beginner at both activities. I am really exciting to try new activities at the Camp. It would be greatful, therefore, if you could advise me what sort of clothes should I take or about money and there are anything that I need for the Camp. I am looking forward to hearing from you.

Yours sincerely

Note: All additional tasks are from the Council of Europe's document "Relating language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment". This can be found online at the following address:

http://www.coe.int/t/dg4/education/elp/elp-reg/Source/Key_reference/exampleswriting_EN.pdf

APPENDIX 4: SPEAKING SCALE

Overall	Descriptor
5 [C]	Consistently high level of grammatical and lexical range and accuracy; errors are rare and difficult to spot. Clear, effective pronunciation and intonation; varies intonation and sentence stress correctly to express finer shades of meaning. Fluent and spontaneously, with little or no sign of effort. Clear, smoothly flowing, well-structured speech, with controlled use of organisational patterns, connectors and cohesive devices.
4 [B2]	Sufficient range and control of grammatical forms and lexis to express ideas without much conspicuous hesitation, using some complex forms to do so. No mistakes lead to misunderstanding. Has clear, effective pronunciation and intonation. Stretches of language with fairly even tempo; can be hesitant when searching for patterns and expressions, fairly long pauses possible. Uses a limited number of cohesive devices to link utterances into clear, coherent discourse; may be some 'jumpiness' in long turns.
3 [B1]	Sufficient range and control of grammatical forms and lexis to get by, but there is hesitation, repetition and difficulty with formulation. A reasonably accurately repertoire of frequently used 'routines', patterns and words associated with more predictable situations, but major errors still occur when expressing more complex thoughts. Pronunciation is intelligible though the accent means that occasional mispronunciations occur. Keeps going comprehensibly; pausing for grammatical and lexical planning and repair is very evident in longer stretches of production. Links a series of shorter, discrete simple elements into a connected, linear sequence of points.
2 [A2]	Control of basic grammatical forms and lexis, but may have to compromise the message and take time to formulate structures. Uses some simple structures and lexis correctly, but still systematically makes basic mistakes (e.g. tends to mix up tenses and forgets to mark agreement; sufficient vocabulary for the expression of basic communicative needs only). Meaning clear. Pronunciation is generally clear enough to be understood despite a noticeable accent and occasional difficulty for the listener. Constructs phrases on familiar topics despite very noticeable hesitation and false starts. Links groups of words with simple connectors like 'and', 'but' and 'because'.
1 [A1]	Very basic range of simple forms with only limited control of a few simple grammatical structures and sentence patterns in a learned repertoire. Basic vocabulary of isolated words and phrases related to particular concrete situations. Pronunciation of a very limited range of words and phrases can be understood with some effort. Manages very short, isolated utterances, with much pausing to search for expressions, to articulate less familiar words, and to repair communication. Little attempt to link words or groups of words, when it happens uses very basic linear connectors like 'and' or 'then'.
0	No or incomprehensible or irrelevant answer

APPENDIX 5: SPEAKING TASKS

Task 1

You will be asked 3 questions. Answer each question as fully as you can. You have a maximum of 30 seconds to answer each question so don't worry if the computer stops you. You will hear this sound (beep) before each question. All your answers will be recorded. The test will now begin. (3 second pause).

Task 2

In this part you will see a picture and answer three questions. Before each question you will hear this sound (beep) You can talk for a maximum of 45 seconds for each question. There are 5 marks for this task.

Task 3

Task 3a

You will see 2 pictures. Look at them and say what you see in the two pictures. You only have 40 seconds to do this. At the end of this time, you will hear a sound (beep). The test will now begin

Task 3b

You should now compare something in these pictures. You have 1 minute for this task. AT the end of 1 minute you will hear a sound (beep). Here is your question.

What would it be like to work in these two places?

Task 3c

This is your last question about the pictures. You have 1 minute to answer. AT the end of this time you will hear this sound (beep). Here is your question.

Which of these places do you think it would be better to work in, and why?

Task 4

Talk about the personal achievement or award you have received.

How did you feel about this achievement?

Do awards encourage people to do their best?

APPENDIX 6: TASK PARAMETERS EXPLAINED

Parameter	Description
Purpose	The requirements of the task. As with tests of other aspects of language ability this gives candidates an opportunity to choose the most appropriate strategies and determine what information they are to target in the text in comprehension activities. Facilitates goal setting and monitoring (key aspects of cognitive validity).
Response format	How candidates are expected to respond to the task (e.g. MCQ, SAF, matching, handwriting, writing on computer etc.). Different formats can impact on performance.
Known criteria	As with listening tests, letting candidates know how their performance will be assessed. Means informing them about rating criteria beforehand (e.g. in SAF, is spelling or grammar relevant as is the case in IELTS; for writing, letting the test takers know about the assessment criteria before they attempt the task).
Weighting	Goal setting can be affected if candidates are informed of differential weighting of items before test performance begins. Items should only be weighted where there is compelling evidence that they are more difficult and/or more central to the domain.
Order of items	In reading comprehension tests, items will not appear in the same order as the information in the text where students search read (i.e. for scanning) but may appear in any order for careful reading.
Time constraints	Can relate either to pre-performance, or during performance. The latter is very important in the testing of reading, as without a time element we cannot test skills such as skimming and scanning (i.e. without this element all reading will be 'careful')
Discourse mode	Includes the categories of genre, rhetorical task and patterns of exposition.
Channel	In terms of input this can be written, visual (photo, artwork, etc), graphical (charts, tables, etc.) or aural (input from examiner, recorded medium, etc). Output depends on the ability being tested.
Text length	Amount of input/output
Writer – reader relationship	This can be an actual or invented relationship. Test takers are likely to react differently to a text where the relative status of the writer is known – or may react in an unpredictable way where there is no attempt to identify a possible relationship (i.e. the test developer cannot predict who the test taker may have in mind as the writer and so the test developer looses a degree of control over the conditions).
Nature of information	The degree of abstractness. Research suggests that more concrete topics/inputs are less difficult to respond to than more abstract ones.
Content knowledge	Same as background knowledge which is very likely to impact on test task/item performance.

Linguistic	
Lexical range	These relate to the language of the input (usually expected to be set at a level below that of the expected output) and to the language of the expected output. Described in terms of a curriculum document or a language framework such as the CEFR.
Structural range	
Functional range	
Physical conditions	All of these elements are taken into consideration in the Information for Centres documents. Centres are routinely monitored to ensure that they are complying with the regulations.
Uniformity of administration	
Security	

Source: O'Sullivan (2009)

BRITISH COUNCIL
APTIS TECHNICAL REPORTS

Linking the Aptis Reporting Scales to the CEFR

Barry O'Sullivan, British Council

www.britishcouncil.org/aptis

ISSN 2057-7168