BRITISH COUNCIL

ENGLISH LANGUAGE
ASSESSMENT RESEARCH GROUP

Technical Report

Aptis General Technical Manual
Version 1.0
TR/2015/005

Barry O'Sullivan, British Council
Jamie Dunlea, British Council

ISSN 2057-7168

www.britishcouncil.org/aptis

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

The authors of this technical manual would like to formally acknowledge the contribution of the members of the Assessment Advisory Board:

- Professor Cyril Weir (Chair)
- Professor Micheline Chalhoub-Deville
- Dr Christine Coombe
- Dr Craig Deville
- Professor Jin Yan.

In addition, the following members of the Assessment Research Group at the British Council contributed to the preparation of the manual:

- Vivien Berry
- Stephen Burrows
- Gwendydd Caudwell
- Judith Fairbairn
- Kevin Rutherford
- John Tucker.

1. INTRODUCTION

1.1 About this manual

This manual describes the content and technical properties of Aptis General, the standard English language assessment product offered within the Aptis test system. The Aptis test system was developed by the British Council, which works directly with organisations to provide tests of English as a Second Language / English as a Foreign Language (ESL/EFL) for a range of assessment needs. The primary audience is test users who need to determine if the test is appropriate to help them make decisions regarding the English language ability of individuals.

This manual provides information on:

- the theoretical framework which has shaped the development of the Aptis test system
- the content of the Aptis General test
- how the Aptis General test is scored
- the technical measurement properties of the Aptis General test, such as reliability.

The manual is also intended to be useful for researchers and language testing specialists who want to examine the validity of the test. It is not intended as a guide to test preparation for test-takers or teachers and trainers preparing others to take the test, although some of the material may be useful for the latter group. Information for these groups is provided separately in the form of a Candidate Guide and other support materials, such as online practice tests.[1]

This manual is divided into four chapters. Chapter 1 is an introduction while Chapter 2 provides an overview of the Aptis test system. Chapter 3 describes Aptis General, divided into four subsections: Section 3.1 gives information on the test users; Section 3.2 describes the test purpose, test structure and content, and test administration; Section 3.3 explains the scoring procedures; and Section 3.4 describes areas for an ongoing research agenda. Chapter 4 provides an overview of the processes of item writing and review, the approach to special accommodations, and an overview of other sources of validity evidence to support the uses and interpretations of Aptis General.

1.2 Intended audience for the manual

Test users, often referred to as stakeholders, include a diverse range of people involved in the process of developing and using a test, and also those who may not be directly involved but are situated within the wider social context in which the test is used and has consequences. This manual is primarily written for a particular group of test users: decision-makers in organisations that are using or considering using Aptis General. A full description of the wider range of various stakeholders and their importance to the process of language test validation can be found in Chalhoub-Deville and O'Sullivan (2015).

Aptis General is used by a wide range of organisations, including educational institutions, ministries of education, and commercial organisations. In the context of how Aptis General is used, decision-makers are those, such as project and department heads, who are tasked with approving the use of a test for their particular needs. Such decisions will often be multi-layered involving participants with different levels of testing expertise, from those with ultimate responsibility for a project who must

[1] http://www.britishcouncil.org/exam/aptis

approve recommendations made by others to those tasked with carrying out the evaluation of available assessment options and making the recommendations to develop or use a particular testing product. Those tasked with making such decisions for particular uses will include training managers and program coordinators for companies and educational institutions, as well as admissions officers in educational institutions and human resources managers in commercial organisations.

The examples given above, while not intended to be exhaustive, make it clear that decision-makers will come from a range of professional experience and backgrounds, and will not necessarily be experts in language assessment. It is important, then, that the review and evaluation of assessment options involves the input of experts on language teaching and assessment who can review the information in this manual to provide expert opinion on the suitability of the test for the uses proposed. While the manual is intended to be as accessible as possible, it is intended to provide the necessary information for making important decisions, and such decisions require an understanding of the relevance of the technical information presented in this manual for the intended uses by the organisation.

1.3 About the British Council

The British Council is the UK's international organisation for cultural relations and educational opportunities. The British Council creates international opportunities for the people of the UK and other countries, and builds trust between them worldwide.

Founded in 1934 and incorporated by Royal Charter in 1940, the British Council is a registered charity in England, Wales and Scotland. We are also a public corporation and a non-departmental public body (NDPB) sponsored by the Foreign and Commonwealth Office.

We are an entrepreneurial public service, earning our own income, as well as receiving grant funding from government. By 2015, over 80 per cent of our total turnover will be self-generated by charging those who are able to pay for our services and expertise, bidding for contracts to deliver programmes for UK and overseas governments, and developing partnerships with private sector organisations. The British Council works in more than 110 countries, and has over 7,000 staff, including 2,000 teachers.

Two of the core aims in the Royal Charter refer to developing a wider knowledge of the English language and promoting the advancement of education. The English language is one of the UK's greatest assets, connecting people around the world and helping to build trust for the UK. We work with UK partners to provide people globally with greater access to the life-changing opportunities that come from learning English and from gaining internationally-respected UK qualifications. We do this through: face-to-face teaching and blended courses; supporting English language teaching and learning in public education systems; providing materials in a wide range of media for self-access learning; and by managing English language examinations and other UK qualifications across the world. Through a combination of our free and paid-for services, and by involving UK providers in meeting the demand for English, we support teachers and learners worldwide.

For more information, visit: www.britishcouncil.org

2. THE APTIS TEST SYSTEM

2.1 Overview

The Aptis test system is an approach to test design and development devised by the British Council for business-to-business (B2B) language assessment solutions. Aptis integrates test design, development, and delivery aspects within an integrated system to provide flexible English language assessment options to test users. The system combines a coherent theoretical approach to language test development and validation with an operational network for content creation and test delivery. Tests are developed within the Aptis system for various uses by different test users, but according to the same theoretical principles of language test validation and the same operational approach to quality assurance. This section of the manual provides a brief overview of the core concepts common to all tests developed within the Aptis system.

2.2 Model of test development and validation

The Aptis test system was based primarily on a test development and validation model advanced by O'Sullivan (2011, 2015a), O'Sullivan and Weir (2011), and Weir (2005). For detailed examples of how the model has been applied in other testing contexts, see Geranpayeh and Taylor (2013), Khalifa and Weir (2009), O'Sullivan and Weir (2011), Shaw and Weir (2007), Taylor (2012), and Wu (2014). As O'Sullivan (2015a) notes: "the real strength of this model of validation is that it comprehensively defines each of its elements with sufficient detail as to make the model operational". Detailed descriptions of these elements can be found in O'Sullivan (2015a).

In practice, the socio-cognitive model is reflected in Aptis in the design of the underlying test and scoring systems. These are operationalised using detailed specifications, again based on the socio-cognitive approach (see Appendices B–F), and supported by exemplar tasks and items (as reflected in the sample tests available on the Aptis website (www.britishcouncil.org/exams/aptis). The specifications demonstrate how tasks are designed to reflect carefully considered models of language progression that incorporate cognitive processing elements explicitly into task design, for example, through the use of the Khalifa & Weir (2009) model for reading, the model suggested by Field (2015) for listening, and the use of language functions derived from the British Council – Equals Core Inventory and the lists for speaking developed by O'Sullivan et al (2002) to form the basis of productive skill tasks. At the same time, detailed attention is paid within the specifications to the contextual parameters of tasks across all components, with the interaction between contextual and cognitive parameters manipulated in explicit ways to derive tasks that are built to reflect specific CEFR levels. The socio-cognitive approach also provides the theoretical foundation for the way in which the concept of localisation is operationalised in Aptis.

The socio-cognitive model has adopted and built on the view of validity as a unitary concept that has become the consensus position in educational measurement following Messick's seminal 1989 paper. This conceptualisation of validity is endorsed by the professional standards and guidelines for best practice in the field (AERA, APA, NCME, 1999; ILTA, 2007; EALTA, 2006). A further important development in validity theory has been the promotion of an argument-based approach to structuring and conceptualising the way the evidence in support of the uses and interpretations of test scores is collected and presented (e.g. Bachman, 2004; Bachman and Palmer, 2010; Chapelle et al, 2008, 2010; Kane, 1992, 2001, 2002, 2013). The conceptualisation of construct and context as presented by Chalhoub-Deville (2003), in which she differentiates between cognitive and socio-cognitive approaches, is also relevant for critically interpreting the model proposed by O'Sullivan (2011), O'Sullivan and Weir (2011) and Weir (2005).

Users of this manual who are interested in situating the model driving the Aptis test system in the wider literature on validation are referred to the overviews of validity theory in O'Sullivan (2011), O'Sullivan and Weir (2011), and Weir (2005). The theoretical discussion is more fully documented and integrated into a critical appraisal of developments in validity theory in the decades following Messick's seminal 1989 paper in Chalhoub-Deville and O'Sullivan (2015).

2.3 Localisation

Localisation is used within the Aptis test system to refer to the ways in which particular test instruments are evaluated and, where it is considered necessary, adapted for use in particular contexts with particular populations to allow for particular decisions to be made.

The following provides a brief description of how localisation is built into the Aptis test system to facilitate a principled approach to the development of variants within the system for particular test uses. The approach described below is operational in focus. It has been derived through consideration of the definition of localisation proposed by O'Sullivan (2011), and informed by the experiences of the Aptis development team in working with test users in diverse contexts. A full discussion of the theoretical underpinning of localisation and a framework for operationalising the concept is available in O'Sullivan and Chalhoub-Deville (2015).

Table 1 identifies five different types of localisation showing the different amounts of adaptation or change that may be required by a particular test user for a particular local context. The Aptis test development team has found it useful to present these different degrees of change in terms of "levels", with a higher level representing a greater degree of change from the standard assessment product. The descriptions in the table presented here are brief, general overviews of key features, and are not intended to be exhaustive or definitive.

The table is intended to provide a general framework to guide the discussion of assessment options for localised needs in a principled way, and to facilitate communication between the Aptis development team and test users by giving broad indications of the degree of time, effort and resources that might be required at each level of localisation.

As noted earlier, Aptis General is the standard assessment option in the Aptis system. Modifications at levels 2 – 4 in Table 1 would generate new variants of Aptis assessment products within the system. Examples of how such a process has worked include Aptis for Teachers (which was developed at a level 2 degree of localisation), and Aptis for Teens (which involved developing new tasks appropriate for learners younger than the typical test users of Aptis General, and thus required a level 4 localisation).

Table 1: Levels of localisation in the Aptis test system

Level	Description	Examples
Level 0	Aptis General (or other existing variant) in a full, four-skills package	User selects a four-skills package of any Aptis (General or variant) available for use.
Level 1	Options for localisation are limited to selection from a fixed range of pre-existing features, such as delivery mode and/or components	User is able to select the skills to be tested and/or the mode of delivery that is appropriate. For example, the Reading package (Core component + Reading component) of Aptis General, taken as a pen-and-paper administration.
Level 2	Contextual localisation: lexical, topical modification	Development of specifications for generating items using existing task formats but with topics, vocabulary, etc. relevant for specific domains (e.g. Aptis for Teachers).
Level 3	Structural reassembly: changing the number of items, proficiency levels targeted, etc., while utilising existing item-bank content.	Developing a test of reading targeted at a specific level, e.g. B1, using existing task types and items of known difficulty calibrated to the Aptis reading scale.
Level 4	Partial re-definition of target construct from existing variants. Will involve developing different task types to elicit different aspects of performance.	Developing new task types that are more relevant for a specific population of test-takers, while remaining within the overall framework of the Aptis test system (e.g. Aptis for Teens).
Level 5	The construct and/or other aspects of the test system are changed to such an extent that the test will no longer be a variant within the system.	For example, developing a matriculation test for uses within a formal secondary educational context; developing a certification test available to individuals rather than organisations, etc.

3. APTIS GENERAL

Aptis General is a test of general English proficiency for adult test-takers. As a business-to-business assessment solution, it is offered directly to institutions and organisations for testing the language proficiency of employees, students, etc. Aptis General is most suitable for situations in which flexibility, efficiency (including cost efficiency), and accessibility are primary concerns.

3.1 Overview of typical test-takers

Aptis General is designed to provide assessment options for ESL/EFL speakers spanning proficiency ranges from A1 to C1 in terms of the Common European Framework of Reference for Languages (CEFR). Test-takers will be 16 years old or older. Learners may be engaged in education, training, employment or other activities.

The description of test-taker variables is necessarily generic for Aptis General, as it is intended to provide cost-effective, flexible testing options which can be made available as ready-to-use products (levels 0–1 of the localisation framework) in a broad range of contexts. Potential test users are expected to engage with the Aptis team to evaluate whether Aptis General is the most appropriate variant for the intended test-taker population.

3.2 Test system

3.2.1 Test purpose

Aptis General is a test of general English proficiency designed for adult learners of English as a Foreign / Second Language (EFL/ESL). The test is provided directly to organisations and is administered at times and locations decided by the test user. The results are intended for use within a particular programme or organisation. The test is not a certificated test and individuals do not apply to take a test directly. Typical uses for which the test is considered appropriate include:

- identifying employees with the language proficiency levels necessary for different roles
- identifying language training needs for employees required to fulfil specific roles
- streaming according to proficiency level within language learning and training programmes
- assessing readiness for taking high-stakes certificated exams or to participate in training programmes
- identifying strengths and weaknesses to inform teaching and support for learners
- evaluating progress within language training programmes.

No specific cultural or first language background is specified in the test design, and test content is developed to be appropriate for learners in a variety of contexts.

The concept of general proficiency, which has underscored the test and task design, was informed through reference to a number of sources, and is described in more detail in O'Sullivan (2015a). The CEFR has been used from the outset to provide a descriptive framework of proficiency to structure the levels targeted and as starting points for task design and content selection. The approach to using the CEFR followed the recommendation of Davidson and Fulcher (2007, p. 232) for test developers to see the framework as a "series of guidelines from which tests…can be built to suit local contextualised needs".

In defining the linguistic parameters of tasks, the British Council – EAQUALS Core Inventory for General English (North, Ortega & Sheehan, 2010) has been used as an important reference point. A further important source of information was the international network of teaching centres operated by the British Council. The development team drew on the assessment needs identified by these centres through working with a diverse range of learners and clients. As outlined in O'Sullivan (2015a), this knowledge and experience was incorporated directly into test and task design through a series of workshops in which British Council teachers and assessment experts, who had participated in a professional development course focused on assessment, worked directly on the design of the test in the development stage.

3.2.2 Target language use (TLU) domain

The test is designed to provide useful feedback on the ability to participate in a wide range of general language use situations in the educational, occupational, and public domains. Potential target language use[2] (TLU) contexts include students in upper secondary (over the age of 16 years), higher education and training programmes, as well as adults using English for work-related purposes. Typical TLU tasks will include those in which learners are using the language to achieve real-world goals, particularly at the intermediate and advanced levels, as well as situations in which language learning itself is the goal of study or training.

Some potential target language use situations include using English:

- to communicate with customers, colleagues and clients
- to participate in English-medium training and education programmes
- in the public domain while travelling for work or study
- to access information and participate in social media and other forms of information exchange online.

In many EFL contexts, learners will have varying degrees of access to authentic input and text outside the training programmes or work environment in which they are being tested. However, English language newspapers, TV and radio programmes, and access to the Internet will provide potential sources of input, particularly for learners at higher (B1+) levels.

[2] For a definition of TLU domain which has been influential in the field of language testing research, see Bachman and Palmer (1996, p. 18).

3.2.3 Test components

The test is primarily a computer-based (non-adaptive) test which can measure all four skills in addition to grammatical and vocabulary knowledge. Tables 2 to 6 present an overview of the structure of the five components which make up the full, four-skills package[3] of Aptis General:

1. Core Grammar and Vocabulary component
2. Listening component
3. Reading component
4. Speaking component
5. Writing component.

As noted in Section 2.3 on localisation, at the 0-level of localisation, an organisation would choose to use the full package with all five components of Aptis General included. The system is designed to promote flexibility by offering organisations the choice, at level 1 of the localisation framework, of choosing which components to include in a package in order to focus resources on those skills most relevant to their needs. The Core component, however, is always included as a compulsory component and used in combination with the other skills as required by the test user.

The Core, Reading and Listening components utilise selected-response formats. Speaking and Writing components require test-takers to provide samples of spoken and written performance. The Speaking test is a semi-direct test in which test-takers record responses to pre-recorded prompts. The task formats across all components make use of the computer delivery mode to utilise a range of response formats, and to approximate real-life language use situations that learners may encounter online (for example, in the Writing component, in which test-takers engage in an online discussion responding to questions). Task parameters such as topic, genre and the intended audience are designed to be relevant to the TLU domain and target test-takers, and are made explicit to help contextualise tasks.

Detailed specifications for each task type used in each component are included in Appendices B to G. Examples of the tasks used in operational tests can be found in the preparation materials provided online, including online practice tests and the Candidate Guide.

[3] The full package option is also referred to as a *four-skills package* because it contains components testing each of the four main skills of listening, reading, speaking and writing in addition to the Core component which tests language knowledge.

Table 2: Overview of the structure of the Core component

Part	Skill focus	Items / part	Lvl	Tasks/ level	Items / task	Task focus	Task description	Response format
1	**Grammar**	**25**	**A1**	5	1	Syntax and word usage	Sentence completion: select the best word to complete a sentence based on syntactic appropriacy.	3-option multiple choice
			A2	5-7	1			
			B1	5-7	1			
			B2	5-7	1			
2	**Vocabulary**	**25**	**A1**	1	5	Synonym (vocabulary breadth)	Word matching: match 2 words which have the same or very similar meanings.	5 target words. Select the best match for each from a bank of 10 options.
			A2	1	5	Meaning in context (vocabulary breadth)	Sentence completion: select the best word to fill a gap in a short sentence. Understanding meaning from context.	5 sentences, each with a 1-word gap. Select the best word to complete each from a bank of 10 options.
			B1	1	5	Meaning in context (vocabulary breadth)	Sentence completion: select the best word to fill a gap in a short sentence. Understanding meaning from context.	5 sentences, each with a 1-word gap. Select the best word to complete each from a bank of 10 options.
				1	5	Definition (vocabulary breadth)	Matching words to definitions.	5 definitions. Select the word defined from a bank of 10 options.
			B2	1	5	Collocation (vocabulary depth)	Word matching; match the word which is most commonly used with a word targeted from the appropriate vocabulary level.	5 target words. Select the best match for each from a bank of 10 options.

Table 3: Overview of the structure of the Reading component

Test	Part	Skill focus	Items/ Part	Lvl	Tasks/ level	Items/ Task	Task focus	Task description	Response format
Reading 25 items	**1**	**Sentence level meaning**	**5**	**A1**	1	5	Sentence level meaning (Careful, local reading)	Gap fill. A short text with 5 gaps. Filling each gap only requires comprehension of the sentence containing the gap. Text-level comprehension is not required.	3-option multiple choice for each gap.
	2	**Inter-sentence cohesion**	**6**	**A2**	1	6	Inter-sentence cohesion (Careful global reading)	Re-order jumbled sentences to form a cohesive text.	Re-order 6 jumbled sentences. All sentences must be used to complete the story.
	3	**Text-level comprehension of short texts**	**7**	**B1**	1	7	Text-level comprehension of short texts (Careful global reading)	Banked gap fill. A short text with 7 gaps. Filling the gaps requires text-level comprehension and reading beyond the sentence containing the gap.	7 gaps in a short text. Select the best word to fill each gap from a bank of 9 options.
	4	**Text-level comprehension of long text**	**7**	**B2**	1	7	Text-level comprehension of longer text (Global reading, both careful and expeditious)	Matching the most appropriate headings to paragraphs. Requires integration of micro- and macro-propositions within and across paragraphs, and comprehension of the discourse structure of more complex and abstract texts.	7 paragraphs forming a long text. Select the most appropriate heading for each paragraph from a bank of 8 options.

Table 4: Overview of the structure of the Listening component

Test	Skill focus	Item/ Part	Lvl	Task/ level	Item/ Task	Format	Task description	Response format
Listening 25 items (The distribution of items across levels is an approximate target and may differ slightly across versions depending on content. The overall difficulty of each test version is constrained to be comparable)	**Lexical recognition**	**10**	**A1**	**10**	1	Monologues	Q&A about listening text. Listen to short monologues (recorded messages) to identify specific pieces of information (numbers, names, places, times, etc.)	4-option multiple choice. Only the target is mentioned in the text.
	Identifying specific, factual information	**5**	**A2**	**5**	1	Monologues & Dialogues	Q&A about listening text. Listen to short monologues and conversations to identify specific pieces of information (numbers, names, places, times, etc.)	4-option multiple choice. Lexical overlap between distractors and words in the input text.
	Identifying specific factual information	**5**	**B1**	**5**	1	Monologues & Dialogues	Q&A about listening text. Listen to short monologues and conversations to identify propositions. The information targeted is concrete and of a factual/literal nature. Requires integration of information over more than one part of the input text.	4-option multiple choice. Distractors should have some overlap with information and ideas in the text. Target and distractors (where possible) are paraphrased.
	Meaning representation / inference	**5**	**B2**	**5**	1	Monologues & Dialogues	Q&A about listening text. Listen to monologues and conversations to identify a speaker's attitude, opinion or intention. The information targeted will require the integration of propositions across the input text to identify the correct answer.	4-option multiple choice. Both target and distractors are (where possible) paraphrased, and distractors refer to important information and concepts in the text that are not possible answers to the question.

Table 5: Overview of the structure of the Speaking component

Test	Part	Skill focus	Lvl	Task description	Channel of input / prompts	Time to plan	Time for response	Rating criteria
Speaking	**1**	**Giving personal information**	**A1/A2**	Candidate responds to 3 questions on personal topics. The candidate records his/her response before the next question is presented.	Questions presented in both written and oral form (pre-recorded). Questions presented in a sequence (e.g. Q2 is presented after the response to Q1).	No	30 seconds to respond to each question	Separate task-based holistic scales are used for each task. Performance descriptors describe the expected performance at each score band. The following aspects of performance are addressed: *1) grammatical range and accuracy* *2) lexical range and accuracy* *3) pronunciation* *4) fluency* *5) cohesion and coherence.*
	2	**Describing, expressing opinions, providing reasons and explanations**	**B1**	The candidate responds to 3 questions. The first asks the candidate to describe a photograph. The next two are on a concrete and familiar topic related to the photo.	1) Questions presented in both written and oral form (pre-recorded). Questions presented in a sequence (e.g. Q2 is presented after the response to Q1). 2) A single photo of a scene related to the topic and familiar to A2/B1 candidates on screen.	No	45 seconds to respond to each question	
	3	**Describing, comparing and contrasting, providing reasons and explanations**	**B1**	The candidate responds to 3 questions / prompts and is asked to describe, contrast and compare two photographs on a topic familiar to B1 candidates. The candidate gives opinions, and provides reasons and explanations.	1) Questions presented in both written and oral form (pre-recorded). Questions presented in a sequence (e.g. Q2 is presented after the response to Q1). 2) Two photographs showing different aspects of a topic are presented on screen.	No	45 seconds to respond to each question	
	4	**Integrating ideas on an abstract topic into a long turn. Giving and justifying opinions, advantages and disadvantages**	**B2**	The candidate plans a longer turn integrating responses to a set of 3 questions related to a more abstract topic. After planning their response, the candidate speaks for two minutes to present a coherent, continuous, long turn.	1) Three questions are presented simultaneously in both written and oral form (pre-recorded). Questions remain on screen throughout the task. 2) One photograph illustrating an element of the topic mentioned in the prompts. The photo is not referred to in the questions.	1 minute	2 minutes for the entire response, integrating the 3 questions into a single long turn	

Table 6: Overview of the structure of the Writing component

Test	Part	Skill focus	Lvl	Task description	Channel of input / prompts	Expected output	Rating criteria
Writing	**1**	**Writing at the word level. Simple personal information on a form.**	**A1**	The candidate completes a form by filling in some basic personal information. All responses are at the word/phrase level, such as name, birthdate, etc.	Form with 9 clearly marked categories (name, date of birth, etc.). There are 9 gaps in the form to be filled.	9 short gaps filled by 1–2 word responses	Separate task-based holistic scales are used for each task. Performance descriptors describe the expected performance at each score band. The following aspects of performance are addressed (not all aspects are assessed for each task): *1) task completion* *2) grammatical range and accuracy* *3) lexical range and accuracy* *4) cohesion and coherence* *5) punctuation and spelling.*
	2	**Short written description of concrete, personal information at the sentence level.**	**A2**	The candidate continues filling in information on a form. The task setting and topic are related to the same purpose as the form used in part 1. The candidate must write a short response using sentence-level writing to provide personal information in response to a single written question.	Written. The rubric presents the context, followed by a short question asking for information from the candidate related to the context.	20–30 words	
	3	**Interactive writing. Responding to a series of written questions with short paragraph-level responses.**	**B1**	The candidate responds interactively to 3 separate questions. Each response requires a short paragraph-level response. The questions are presented as if the candidate is writing on an internet forum or social network site. The task setting and topic are related to the same purpose/ activity used in parts 1 and 2.	Written. The rubric presents the context (discussion forum, social media, etc.). Each question is displayed in a sequence following the completion of the response to the previous question.	30–40 words in response to each question	
	4	**Integrated writing task requiring longer paragraph-level writing in response to two emails. Use of both formal/ informal registers required.**	**B2**	The candidate writes two emails in response to a short letter/notice connected to the same setting used in parts 1, 2 and 3. The first email is an informal email to a friend regarding the information in the task prompt. The second is a formal email to an unknown reader connected to the prompt (management, customer services, etc.)	Written. The rubric presents the context (a short letter/ notice/ memo). Each email is preceded by a short rubric explaining the intended reader and purpose of the email.	First email: 40–50 words Second email: 120–150 words	

3.2.4 Mode of delivery

Aptis General is usually taken as a computer-based test (CBT). The CBT system uses the Internet to download tests and upload the responses of test-takers to a secure server. While the test-taker interacts directly with the test delivery interface, the system also integrates item production and item banking, the creation of new test forms from the item bank, the administrative elements of registering and scheduling test-takers, the marking of productive skills by human raters, and the reporting of results to the test administrators in charge of test use for a particular organisation.

Multiple versions of each component are made available for live administration at any one time. All versions are created to the same rigorous specifications and undergo the same standardised quality assurance and analysis procedures to ensure comparability (see Sections 3.3.2.1 and 3.3.3.5 for an overview of the approach to maintaining comparability across versions). Within the CBT delivery mode, versions available for live administration are randomly allocated to candidates to enhance security. The system is designed to prevent the same live version of a component being presented to the same candidate twice when the same candidate (registered once with the same details) is scheduled to take the test more than once.

At the same time, in accord with the intention to provide flexible assessment options for organisations with different needs and contexts of use, other delivery mode options are also available. The Core, Reading, Listening, and Writing components can be administered in pen and paper formats, and the Listening and Speaking components are available through a telephone delivery option. The CBT test is also available for administration on tablets. The structure of the tests in terms of components, task types and number of items is the same across delivery modes. While the various delivery modes are offered to provide flexible options, the CBT format is at the core of the system, and as such, there are differences in the number of test forms available for use in different modes, certain modes will entail longer time schedules for the delivery of results than the default CBT mode, and different procedures will be required to ensure fair and secure administration.

Potential test users will need to engage in a discussion with the Aptis team to consider the best delivery mode options for their particular testing context and needs.

3.2.5 Administration and security

Aptis General is sold directly to organisations, not individually to test-takers. Times and locations for administration of the test to the employees, students, etc., in an organisation using the test are agreed between the organisation and the British Council. Organisations have the option of requesting the British Council to perform test set-up and invigilation functions directly or of carrying them out themselves. Tests are generally administered on the organisation's premises, using computer facilities arranged by the organisation. In such cases, test administration, invigilation, and test security will generally be the responsibility of the organisation.

The British Council prepares detailed guides which clearly describe all aspects of the administration of the test, from seating arrangements to the technical requirements for microphones and speakers necessary to deliver speaking and listening tests. Organisations use Aptis General for a range of purposes, and the degree of security required for fair administration and consistent interpretation of results will differ accordingly. As such, the individual needs of an organisation and the intended use of the test are discussed directly with the British Council. Guidelines appropriate for each organisation are then developed in consultation with the British Council.

Organisations have the option of being set up as a virtual test centre for the purposes of administering the test through the CBT system, or requesting an existing British Council centre to carry out those administrative functions. Administrators associated with a test centre that is registered in the system have the ability to register test-takers, schedule tests, monitor the progress of tests that have been scheduled and access results for test-takers once the tests have been completed and results finalised within the system.

Test security is the joint responsibility of the test user and the British Council. The security of the test system and the test content is managed through the computer delivery system by the British Council, which oversees the creation of test content from item writing through pre-testing and the creation of live test forms, as well as the marking and finalisation of all results. However, the set-up and administration of tests, including the invigilation of test-takers during the test, is often managed directly by the organisation using the test. This system provides organisations with cost-effective, flexible options for administration. The responsibilities of organisations in terms of ensuring fair and secure testing appropriate to their intended uses of the test are stressed clearly to all test users. This joint responsibility is a key feature of the testing program, and is closely linked to the appropriate use and interpretation of Aptis General test results. Aptis General is used within organisations and is not a certificated test (i.e. does not provide proficiency certification which can be used across organisations or contexts outside the original context of testing) partly because the security and integrity of administration is integrally connected to, and determined by, each organisation using the test.

3.3 Scoring

3.3.1 Overview of scoring and feedback

The Core, Reading and Listening components are scored automatically within the computer delivery system. This ensures that accurate results are available immediately following testing. Trained human raters mark the Speaking and Writing components, using an online rating system. A skills profile is provided which reports both a scale score (between 0 and 50) and a CEFR level for each of the four skill components. A CEFR level is not reported for the Grammar and Vocabulary component. As noted in Section 3.2.1, the CEFR has been incorporated into the task and test design for Aptis General from the development stage. The link to the CEFR was further validated through an extensive standard-setting study to set cut-off scores marking the boundary between CEFR levels on the Aptis score scales (O'Sullivan, 2015b).

Table 7 shows the levels of the CEFR with the accompanying designation used for reporting in Aptis General. The level description column contains the level description used in the CEFR. The levels highlighted in yellow indicate those levels at which tasks in Aptis General are specifically targeted: A1 to B2 (for features of tasks at each particular level of the CEFR targeted, see the task specifications in the appendices). If a candidate does not receive a high enough score to be awarded a CEFR level, then they will receive an A0 level (sometimes referred to as pre-A1 or pre-beginner). On the other hand, a candidate who receives a near perfect score will receive a level classification of C. This means the candidate has demonstrated a strong performance at the levels targeted by Aptis and is likely to be able to deal with tasks at the next highest level beyond B2. Aptis General does not distinguish between C1 and C2. The threshold at which a candidate could be considered to have demonstrated a strong enough performance to be classified as being more likely to belong to the next highest CEFR level beyond B2 was investigated during the comprehensive standard-setting study undertaken to set cut-offs for each level on each of the four skill components (O'Sullivan, 2015b). For each of the skills, participants in the standard-setting panels were asked to identify the threshold marking the boundary between B2 and C using the same methodology and approach as was used for identifying the boundaries between the other levels (O'Sullivan, 2015b).

Table 7: CEFR levels reported by Aptis General

Level description in CEFR	Levels in CEFR	Levels reported in Aptis General
Proficient User	C2	C
	C1	
Independent User	B2	B2
	B1	B1
Basic User	A2	A2
	A1	A1
		A0

The cut-off scores for CEFR level designations have been set separately on the scale for each skill component. As the scale and CEFR cut-off scores are distinct for each skill component, scale scores should not be compared directly across skills. A scale score of 30 on one skill (e.g. Reading) should not be interpreted as having the same amount of ability or being at the same CEFR level as a scale score of 30 on a different skill. Scores and CEFR level designations within the same skill are comparable across different versions of the same component and across different administrations of the test. (See Sections 3.3.2.1 and 3.3.3.5 for a description of the approach to maintaining comparability across versions of each component.)

3.3.2 Reliability of receptive skill components

In practical terms, reliability refers to "the consistency of the test results, to what extent they are generalisable and therefore comparable across time and across settings" (ILTA, 2007). All tests contain some degree of measurement error (APA/AERA/NCME, 1999; Bachman, 2004; Weir, 2005). It is thus an important responsibility of test developers to report estimates of the reliability of a test (e.g. APA/AERA/NCME, 1999; ILTA, 2007).

Bachman (2004, p. 160) notes four sources of measurement error associated with inconsistent measurement: 1) internal inconsistencies among items or tasks within the test; 2) inconsistencies over time; 3) inconsistencies across different forms of the test; and 4) inconsistencies within and across raters. The four main types of reliability described in the 1999 Standards for Educational and Psychological Measurement (AERA, APA, NCME) address these sources of error: internal consistency estimates of reliability, test–retest estimates of reliability, parallel forms estimates of reliability, and inter- and intra-rater estimates of reliability. Various methods of estimating the degree to which test scores are free of error associated with these potential sources have been devised to provide indices of reliability generally measured on a scale of 0 to 1, with 1 representing a perfectly reliable test. As noted above, in practice, no test is completely free of measurement error, but the higher a reliability coefficient is, the more confidence test users can have in the results provided by the test.

Bachman (1990, p. 184) suggests that internal consistency should be investigated first since "if a test is not reliable in this respect, it is not likely to be equivalent to other forms or stable across time". At the same time, Weir, (2005, p. 31) notes that "the use of internal consistency coefficients to estimate the reliability of objectively scored formats is most common and to some extent, this is taken as the industry standard". The following section provides estimates of the internal consistency reliability for the Core (grammar and vocabulary), Reading and Listening components of Aptis General. Estimates of rater reliability for the productive skills components are discussed in Section 3.3.3.5.

For a more detailed discussion of reliability specifically in relation to language testing, including formulas for calculating the different kinds of reliability coefficients discussed above and overviews of the limitations and caveats associated with them, see Bachman (1990, 2004) and Weir (2005).

The following internal consistency reliability estimates were derived using operational test data from all versions of Aptis General delivered through the CBT mode in live administrations between April and September 2014. As noted in Section 3.2.3, test users may select different combinations of skills components, e.g. some candidates taking a full package with all five components, while others may take only a Reading package (with the Core and Reading components) or some other combination. As such, there are different numbers of candidates in the data set for each component. The reliability indices were calculated separately for each version in each component using the Kuder-Richardson 21 formula[4]. Table 8 gives an overview of the sample sizes used in the analysis for each component, noting the average number of candidate scores used in each version, the maximum and minimum number of candidates on any version, and the total number of candidate scores available across all versions for each component. Table 9 shows the average, maximum and minimum internal consistency reliability estimates across the versions of each component in the analysis.

Table 8: Overview of sample sizes used in estimation of reliability

	Mean	Max	Min	Total
Core (G&V)	2145	2190	2099	15014
Listening	1408	1438	1381	9857
Reading	1721	1757	1690	12048

Table 9: Reliability estimates across operational versions of Aptis General

	Core (G&V)	Listening	Reading
Mean	0.91	0.82	0.89
Max	0.93	0.85	0.91
Min	0.88	0.79	0.85

[4] KR-21 is a shortcut estimate of KR-20, which is a special case of Cronbach's alpha for dichotomous items (Bachman, 2004, p. 163). The formula for KR-21 requires only the mean and variance of the total scores. KR-21 will generally be slightly lower than KR-20 or Cronbach's alpha, which are considered to be lower bounds of internal consistency reliability estimates (Bachman, 2004, pp. 163–166). The estimates shown here are conservative estimates of the internal consistency reliability for live versions of the receptive skills components of Aptis General.

In interpreting reliability estimates, Fulcher and Davidson (2007, p. 107) suggest 0.7 as a minimum requirement, while "high-stakes tests are generally expected to have reliability estimates in excess of 0.8 or even 0.9". The estimates shown in Table 9 demonstrate levels of reliability appropriate for the proposed uses and interpretations of Aptis General, and are generally consistent with figures reported in the literature for large-scale, standardised language proficiency tests, including those used in high-stakes situations (see for example, Chapelle et al, 2010; Weir, 2005; Weir and Milanovic, 2003).

3.3.2.1 Pre-testing and equating for receptive skills components

All items for receptive skills components which employ selected response item and task formats are pre-tested on representative samples of test-takers typical of the variant of Aptis for which the items will be used. The minimum sample size for pre-testing is 100 test-takers. Test-takers are recruited through British Council test and teaching centres internationally. Each sample of 100 (or more) test-takers will be drawn from at least two different geographical and cultural contexts.

At the pre-testing stage, new items created by trained item writers according to test task specifications are mixed with anchor items (see Section 4.1.2 for a description of the item production process). Anchor items are items for which the technical properties, including empirical difficulty are known. The anchor items have difficulty estimates derived on what is known as a logit scale through Rasch analysis. Rasch analysis is one of a family of Item Response Theory models used in educational measurement. Rasch analysis enables the estimation of item difficulty and test-taker ability on a common scale of measurement (Bachman, 2004). Anchor items used in pre-testing have difficulty estimates derived during the field testing of the first version of the first variant of Aptis. The anchor items thus allow all new items to be analysed within the same common frame of reference as the first version of the first variant of Aptis. This version is thus the base or reference version for a common Aptis measurement scale. New test items are placed on the same common scale of measurement through a process known as equating, which is facilitated by the use of the anchor items.

During pre-testing, items are analysed for both empirical difficulty and technical quality in terms of discrimination. Items that meet pre-set quality control criteria are stored in an item bank for use in future operational tests.

3.3.3 Reliability of productive skill components

3.3.3.1 The rating system

Aptis General uses a secure online rating system that allows raters with appropriate authorisation to rate test-taker responses remotely. Raters can be recruited and trained, and then carry out rating wherever they are located, provided they have sufficient Internet access and computer facilities. This functionality greatly enhances the flexibility of the rating system, and extends the reach of the potential rater pool. The system has several advantages. Firstly, it enhances one of the primary goals of the Aptis test system, namely providing efficient and flexible assessment options for organisations. Having raters based in various locations internationally ensures that responses can be rated rapidly regardless of the time zone in which a particular test has been taken. From the perspective of ensuring quality, the system allows for various features for quality control to be integrated into the system, which would be difficult to include in more traditional rating scenarios. The Examiner Network Manager, along with a team of senior raters, monitor all rating through the online system, allowing them to review the status of test-taker responses that have been uploaded to the system, and to constantly monitor the performance of raters.

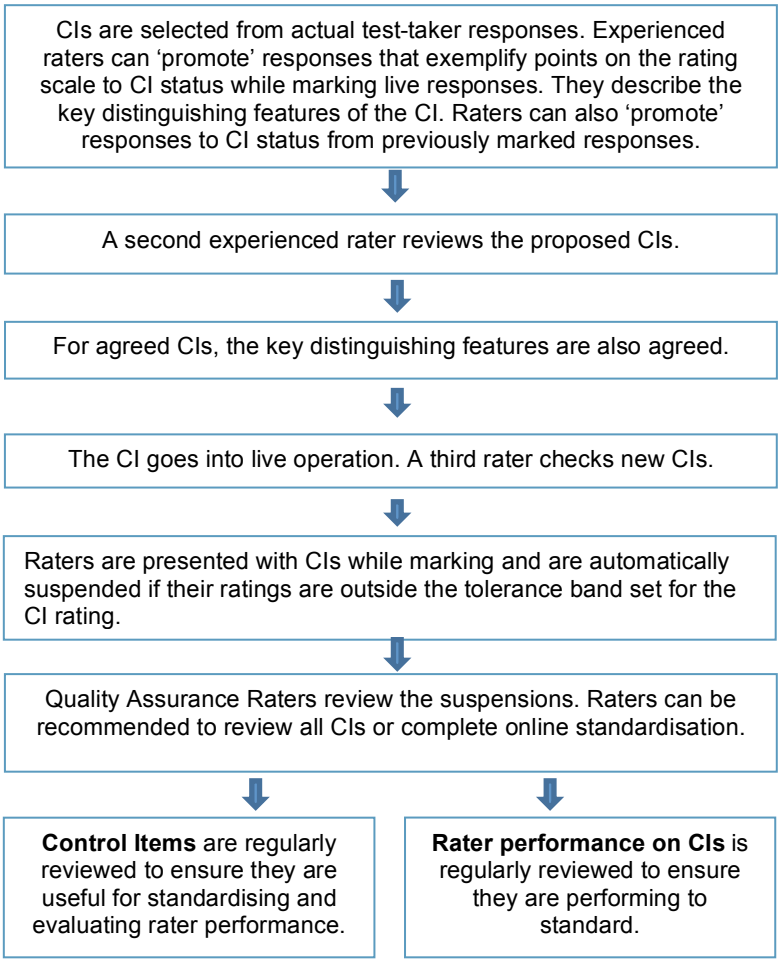
The online rating system automatically breaks up a test-taker's performance on a full Speaking or Writing test into the separate responses for each task (see Table 5 and Table 6 for an overview of the tasks in each component). The same rater will not be able to rate more than one task performance for the same test-taker. This ensures that every test-taker's complete performance across all tasks in a productive skills component is rated by multiple raters. Raters see no information which can identify a candidate or the responses associated with any particular candidate, and they do not have access to the scores given by other raters for performances by the same candidate on other tasks. This ensures the complete security and impartiality of the rating process.

While the complete test performance is thus rated by multiple raters (four raters, one for each task), each specific task performance is single rated. The decision to employ single rating of each task performance was taken to achieve the best possible balance between the demands for fast, cost-efficient assessment services required by organisations and businesses, and the need for valid and reliable scoring that is fair to test-takers and provides test users with the most useful information for the decisions they need to make.

The rating system for Aptis General makes full use of the functionality of the online rating system to implement checks and balances to ensure the technical quality of the scores awarded. In addition to the system described above, to ensure that a test-taker's total score on a productive skill component is derived from scores from multiple raters (across tasks), an ongoing quality-control monitoring system, described below, is integrated within the system to ensure raters are marking to standard.

The online system allows for a comprehensive quality control process to be integrated into the rating procedure by placing pre-scored performances in the responses to be rated by each examiner. This approach has been described by Shaw and Weir (2007, p. 307) as "gold standard seeding". Within the Aptis test system, these pre-scored benchmark, or gold standard, performances are referred to as control items (CIs). Raters are aware that they will be presented with CIs, but there is no distinction in presentation between CIs and operational responses for live marking. When raters begin marking a task type for a particular version of the Speaking or Writing component, they will be presented with a CI for that task type for that version. If the rater awards a score outside of the tolerance band for the pre-agreed score for the CI, then that marker is automatically suspended from rating that task. Once an examiner begins marking live responses, approximately five per cent of performances rated will be CIs. Figure 1 has been adapted from Fairbairn (2015) to provide an overview how the CI system works in practice.

Figure 1: Overview of control item (CI) system (from Fairbairn, 2015)



3.3.3.2 Rater training

All raters are trained using a standardised system. Raters are also expected to pass an accreditation test at the end of the training event. Rater training is carried out using an online training system. The online training system has the same advantage as the online rating system in that it allows for a very large pool of potential raters, and facilitates cost-effective, efficient training as raters can undertake training where they are based without travelling to a face-to-face training event. During training, raters interact directly through discussion forums, etc., with all of the raters in the training cohort and the facilitators supervising the training (the Examiner Network Manager and/or senior examiners).

Raters are given familiarisation training on the CEFR, as the CEFR forms an important part of the rating scale and task design. They are trained in the use of the rating scales developed specifically for the Aptis General productive skills components. During training, they rate a number of standardised, benchmarked examples of performance, receiving feedback from the training facilitator, as well as carrying out discussion with other trainees. Following accreditation and operational rating, in-service training is also provided for raters who do not meet the required level of accuracy or consistency.
A research study investigating the effectiveness of the online training in comparison with face-to-face training (Knoch and Fairbairn, 2015) has been conducted and recommendations from that study are being incorporated into the training program.

3.3.3.3 Rating scales

The rating criteria for both the Speaking and Writing components are based on the same socio-cognitive framework of language test development and validation that underpins the tasks used to elicit performances. The rating criteria, as with the task specifications, are closely linked to the CEFR. Descriptors used within the rating scales are designed to target the kind of performance described within the CEFR. Task specific scales have been developed for each of the tasks in the Speaking and Writing components. The scales are shown in Appendix H. The current rating scales were introduced for operational use in December 2014 following a comprehensive scale revision and validation project (Dunlea and Fairbairn, 2015).

Tasks 1 to 3 for both Speaking and Writing components are rated on scales ranging from 0–5, while Task 4 for both components is rated on a 0–6 scale. Descriptors are provided to describe performance at each score point on the rating scale for that task. The 3 and 4 point score bands describe the target-level performance for a task. For example, Task 3 for Writing is targeted at a B1-level of performance, and the 3 and 4 point score bands describe performance appropriate for a B1-level candidate. The 1 and 2 point bands describe performance on that task which is below the target level. For Task 3, which is targeted at B1, the 1 and 2 point score bands describe performances which would be at the A2 level. The 5 point score band is allocated to performances that are beyond the target level. The ratings provided by raters on the 0–5 or 0–6 scales are subsequently weighted automatically within the system so that tasks targeted at a higher level are weighted more than tasks targeted at a lower level (e.g., for Writing, a high target level performance of 4 on the B2-level task is weighted higher than a high target level performance of 4 on the B1-level task, and so on).

3.3.3.4 Inter-rater reliability

The inclusion of CIs in the online rating system can be used to provide operational estimates of rater reliability. Correlations between raters and their first attempts at CIs can be calculated as a means of estimating the degree of consistency between raters and the intended benchmark scores for CIs. Inter-rater and intra-rater reliability can also be calculated using correlations between all pairs of raters who have marked the same CIs, and between the same rater's marks on the same CIs over time. The following section provides an outline of a pilot study on inter-rater reliability utilising CI data carried out by Fairbairn (2015).

The pilot study examined the scores awarded on CIs for Task 4 for both Speaking and Writing between January and March 2015, the first full three months of operational use of the revised rating scales. As raters may be presented with the same CI multiple times in the course of operational rating, only the first attempt at a CI was used. As all Task 4 responses are rated using the same rating scale, the raters' scores on their first attempt for all CIs on Task 4 across all operational versions of a component were combined into a single column for each rater. The data file thus included multiple columns, one for each rater and also a column for the benchmark CI score, and multiple rows of data, one for each CI performance. A total of 38 CIs for Speaking and 35 for Writing were used in the analysis. Only raters who had scores on a minimum of 15 CIs were included, which resulted in a final data set of 17 raters for Writing and 23 for Speaking. A Pearson product moment correlation matrix was generated for the data set. When averaging multiple correlation coefficients, it is recommended to use a Fisher Z transformation to account for the inherent distortion in correlation coefficients (Bachman, 2004; Hatch and Lazaraton, 1991). This procedure was followed and the average of the transformed correlations was then converted back to the correlation metric. The mean correlations between all pairs of raters on CIs for Task 4 for both Speaking and Writing, and the mean correlations between raters and the benchmark CI scores for the same CIs are reported in Table 10. As with the reliability indices for receptive skills reported in Section 3.3.2, these figures indicate high levels of inter-rater reliability (see for example, Chapelle et al, 2010; Weir, 2005; Weir and Milanovic, 2003).

These figures need to be interpreted in context, however, and are presented only as one form of evidence to help test users to evaluate the scoring validity of the Aptis General productive skills components. The figures shown here were based on one pilot study utilising performances selected for use as Control Items. CIs are selected on the basis of being very clear examples of the performances characterising each score band. The inter-rater correlations generated by this study are thus likely higher than the correlations that would be seen for ratings based on a sample of performances which included more borderline and problematic examples. While this study has important limitations, the use of CI data to investigate inter-rater reliability is an innovative way to obtain rating data from multiple raters on the same items under operational rating conditions. Because of the nature and demands of scoring operational tests, particularly in single rating designs, it is often not possible to obtain such data except through specially designed rater reliability studies conducted outside the operational testing environment. The approach taken here thus offers a way to gain insights into rater consistency under operational conditions, but needs to be followed up with further studies, including specially designed multiple-rating studies carried out outside the normal operational rating environment. Other measures of rating quality will also be addressed in the future, for example through the use of multi-facet Rasch model (MFRM) analysis.

Table 10: Mean correlations on Task 4 CIs for Writing and Speaking

Component	All pairs of raters	Raters with CI benchmark
Speaking	.89	.94
Writing	.97	.97

3.3.3.5 Ensuring comparability in productive skills components

Comparability for different forms of productive skills components is maintained through a combination of rigorous test specifications for item writers, the use of explicit rating scales which have undergone validation, and standardised training of raters to ensure the consistent application of the rating criteria to task performances. This approach is consistent with that employed in most large-scale, standardised testing programs with productive skills components.

As with many such large-scale, standardised tests, new versions of productive skills components are not pre-tested with large groups of test-takers in the same way as they are for receptive skills. Pre-testing for productive skills components is problematic for several reasons, including protecting the security of the test items and the difficulty of using typical equating techniques due to the small number of items that can typically be used for productive skills.

A comprehensive system of quality control and review is carried on new versions for productive skills components to ensure the content of all new versions complies strictly with the task specifications. Ongoing qualitative information is also obtained from raters to inform the periodic operational review of quantitative data to evaluate the performance of test versions over time.

3.3.4 Precision of scoring: Standard Error of Measurement

As noted in Section 3.3.2, all tests contain a certain amount of measurement error. Reliability estimates provide an estimate of the consistency of measurement of the test scores for a specified population of test-takers, but these estimates do not give us a direct indication of the impact of the degree of inconsistency (or measurement error) on an individual's test result (Bachman, 1990; Bachman, 2004; Weir, 2005). A measure useful for interpreting the accuracy of individual scores is the Standard Error of Measurement (SEM), which is calculated according to the following Formula 4.1 (from Bachman, 2004, p. 173).

$$SEM = S_x\sqrt{1 - r_{xx'}}$$

S_x is the standard deviation of the scores and

$r_{xx'}$ is a reliability estimate for the test scores (e.g. KR-21, inter-rater reliability)

The SEM is used to provide an indication of how confident we are that the score obtained by a test-taker on a particular administration of the test reflects his or her "true score" (Bachman, 1990; Bachman, 2004; Weir, 2005). The SEM is reported on the same score scale as the test, so the SEM helps us to understand how large the test error is. The smaller the number for the SEM, the more accurate the test will be. A test-taker's true score, which can never be measured without a perfect test free of error, is likely to fall within a defined range around their observed score. The SEM provides an estimate of that range. If a test-taker were to take a test again, the score obtained would be 68 per cent likely to fall within +/- 1 SEM of their observed score. Table 11 provides estimates of the average SEM for operational versions for each of the five components of Aptis General.[5]

Table 11: Estimates of Standard Error of Measurement (SEM) for Aptis General components

	Core G&V	Listening	Reading	Speaking	Writing
Scale score	0–50	0–50	0–50	0–50	0–50
SEM	3.2	4.5	3.8	3.7	2.0

3.3.5 Using the CEFR in score reporting

The CEFR has been incorporated into the Aptis system from the design and development stage. From that perspective, the functional descriptors of language proficiency contained in the Illustrative Scales of the CEFR have been incorporated into the design and validation of tasks.

The link with the CEFR has further been validated through a standard-setting study carried out in accordance with procedures outlined in the manual produced by the Council of Europe (2009) and updated by O'Sullivan in the City and Guilds 'Communicator' linking project (2009, 2011b). Details of the standard-setting study are reported in a separate technical report (O'Sullivan, 2015b).

[5] SEM for the Core, Listening and Reading components was calculated using the standard deviation of scale scores for live versions in the same operational data used for the analysis of internal consistency in Section 3.3.2, and the KR-21 estimate for each version was used as the reliability estimate. For Speaking and Writing, the analysis used the standard deviation of scale scores for live versions from the same period as the study reported in Section 3.3.4. The inter-rater reliability estimates in Table 11 were used as the reliability estimates.

The study findings can be summarised as follows:

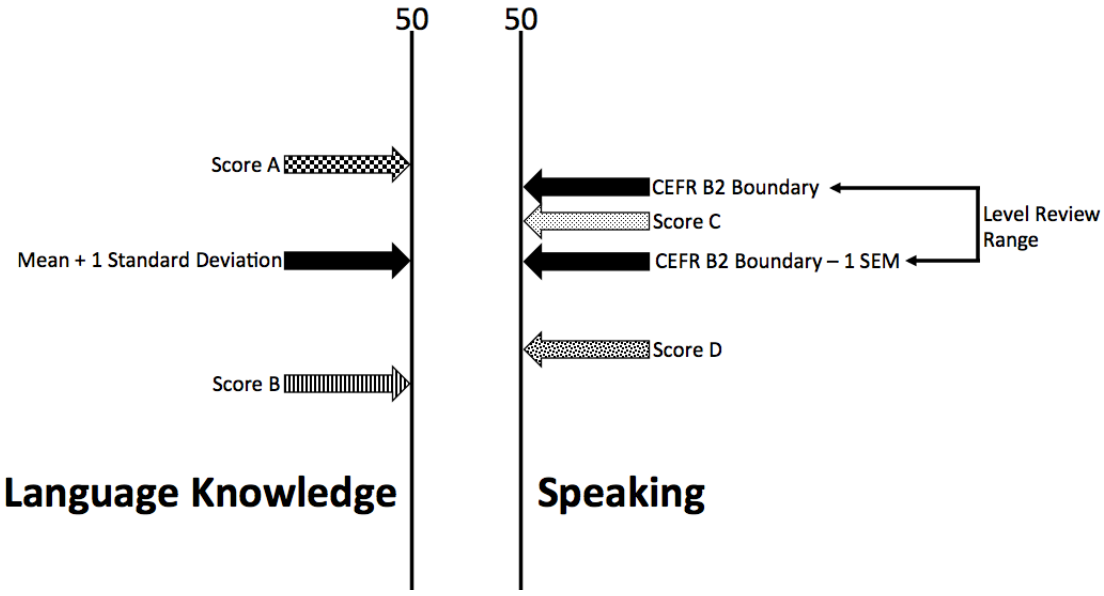
1. The Aptis components in the main variant of Aptis offer a broad measure of ability across the different skills, as well as the key area of knowledge of the system of the language.
2. The Aptis components in the main variant of Aptis are robust in terms of quality of content and accuracy and consistency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

3.3.5.1 Incorporating SEM into the allocation of CEFR levels

Where a candidate achieves a score on one of the main skills components that falls within 1 standard error of measurement (SEM) of a CEFR level boundary, then their score on the Core language knowledge component is taken into consideration when deciding whether they should remain at the lower CEFR level or whether they should be upgraded to the higher level. To receive this upgrade, they should perform significantly above the average on the Core component (set as 1 standard deviation above the mean). This system is intended to increase the accuracy of the CEFR level decisions and contributes significantly to the increased reliability of the outcomes.

In the example shown in Figure 2, a candidate who achieves Score A on the Core component, which is clearly above the review point (Mean plus 1 standard deviation), will have his or her Speaking score adjusted automatically by the system. If, like Score C, it falls within the level review range (boundary point minus 1 SEM), then the person 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 the Core component (Score B), then no action is taken regarding the Speaking score, regardless of where the Speaking score lies in relation to the level review range. This review and adjustment is undertaken automatically within the system. The reported scores on the scale of 0–50 for test-takers are not adjusted, only the CEFR level to which the test-taker will be allocated.

Figure 2: Example of how Core component score is used



The role of the Core component in being a strong predictor of performance on the four skills components is demonstrated by the correlation matrix shown in Table 12. A subset of operational data from the data used for the reliability analysis in Section 3.3.4 was used to generate a Pearson product moment correlation data matrix between the five components. Scale scores from 6,101 test-takers who had taken a complete package with all five components were used to generate the correlation matrix. As can be seen, there are moderate to high correlations between all skills, and the highest correlation for all skills is with the Core component.

Table 12: Correlations between total scores on Aptis General components

	Core	Reading	Listening	Writing	Speaking
Core	1				
Reading	0.75	1			
Listening	0.72	0.68	1		
Writing	0.73	0.71	0.62	1	
Speaking	0.68	0.63	0.65	0.66	1

3.3.5.2 Why CEFR levels are not reported for the Core component

The Core grammar and vocabulary component is central to the design of Aptis for two reasons. Firstly, because of the importance of grammar and vocabulary knowledge as a foundation for the four main skill components reported by Aptis: Listening, Reading, Speaking and Writing. Secondly, in terms of test scores, research has consistently shown grammar and vocabulary to be strong predictors of L2 proficiency (see for example, Shiotsu, 2010; Milton, 2013; van Zeeland and Schmitt, 2012). The grammar and vocabulary component has been positioned as the Core component to enable reference to this stable, valuable predictor of performance for purposes of comparisons across samples and within samples, and also to aid in clarifying borderline scores, enabling more robust reporting of CEFR levels for the four main skills packages.

CEFR levels are not reported for the Core component at this stage because the position of grammar and vocabulary knowledge within the CEFR is one of the most under-specified elements of the framework. Scales for linguistic range, vocabulary range and control, and grammatical accuracy are provided in the CEFR. However, as the CEFR is designed as a multilingual framework general enough to be relevant to a range of languages, the descriptors by design do not contain detailed language-specific information or lists of grammatical or vocabulary items at each level (Council of Europe, 2001, p. 30). Users of the CEFR are encouraged to consider their own contexts and develop detailed language specifications appropriate for those contexts (Council of Europe, 2001, p. 33). Research is ongoing to clarify the relationship between the Core component and CEFR levels.

3.3.5.3 Reporting overall CEFR levels

Overall CEFR levels are reported as a standard element of the Aptis General reporting structure to provide an extra layer of feedback for test users. Overall CEFR levels are calculated by averaging the CEFR levels achieved across all four skill components. An overall CEFR level is only generated when a full package (all five components) is taken. When an overall CEFR level is reported, test users are encouraged to examine the profile of CEFR levels across skills in addition to the overall level. Many learners are likely to have varying abilities across the four major skills. For this reason, for instruction, training, or any other substantive use, it is important to use the valuable information that Aptis reports by looking at a candidate's proficiency profile, in addition to the overall CEFR level.

3.4 The need for ongoing research

The data relating to scoring validity offered in this manual should not be interpreted as static or definitive. It is intended as the first step in an ongoing research agenda to build a robust body of evidence on the technical properties of the test for test users. Quality assurance is an ongoing process. Data collection and analysis in relation to the scoring system needs to be carried out regularly on operational data, as well as through specially designed studies which enable the collection of data and the use of analysis techniques which might not always be possible under operational conditions. Accordingly, this Technical Manual has been labelled as Version 1 to recognise the intention to periodically update the manual with new and revised statistical information.

Future versions will need to address issues of stability of the analysis framework and item bank over time, looking at, for example, the impact of anchor drift on item estimation stability during the pre-testing phase, and investigating the stability of item difficulty measures using larger operational data sets. The reliability statistics reported in this version of the manual will need to be bolstered by Rasch reliability estimates, estimates of decision consistency and reliability appropriate for use with criterion-referenced tests that set grade-level cutoffs – as with the CEFR levels reported by Aptis – and indicators of item performance, such as Rasch-based fit indices and classical testing theory discrimination indices. Similarly, in the investigation of scoring validity for the productive skills, future versions of the manual should report on investigations of rater drift and, as noted in Section 3.3.3.4, extend the range of analysis techniques employed to include MFRM analysis and rating data obtained from a wider range of performances than is possible through using the CI system.

4. Other documentation

4.1 Description of the test production process

4.1.1 Distinguishing between development and production cycles

The description of the test production cycle below describes the ongoing creation of tasks and live test versions for an existing test variant within the Aptis test system, Aptis General. Prior to reaching the stage at which test and task specifications are available to guide the generation of multiple versions of a test which can be treated as comparable and interchangeable, a comprehensive test development process is followed for the design and validation of those specifications. The development cycle for Aptis General is explained in outline in O'Sullivan (2015a). Once a new variant has been through that development process, including large-scale field trialling and statistical analysis, the focus turns to ensuring the ongoing production of multiple versions that are comparable in terms of difficulty and test content. The following sections describe that process of ongoing production of live versions for Aptis General.

As noted in Section 3.2.4, an integrated CBT delivery system is at the core of the Aptis General test. While initial stages of the item production cycle take place outside this system, the majority of the item authoring and test construction stages take place within the system. Central to all stages of task and test construction are the specifications. All individual test tasks are constructed according to rigorous task specifications (see Appendices B to F), which ensures that individual tasks targeted at the same level and designed to measure the same abilities are comparable. Test specifications (see Tables 2 to 6) provide the design template for creating new versions of each test component, ensuring the construction of these versions is consistent and versions are comparable in terms of content and difficulty. Quality assurance, pre-testing, and analysis and review stages are integrated into the production cycle to further ensure this comparability.

4.1.2 The production cycle

Appendix J provides a graphical depiction of the test production cycle from the point of commissioning new items and tasks to the point of final construction of test versions for operational use in live tests. Appendix J presents this cycle as a flow chart, depicting the various points at which different members of the test production team interact with the items and item writers, including the review, revision, and pre-testing of items, as well as the provision of feedback to item writers. The various stages of this cycle are explained in more detail below.

4.1.2.1 The commissioning process

Only trained item writers are asked to submit items for use in the test production process (see Section 4.1.2.5 for a description of the training procedures). Item writers indicate their availability for item writing work over a calendar year, and they are offered commissions on this basis. For any given commission, an item writer is sent an email with the proposed number of items and the deadline for delivery and the item writer confirms acceptance of the commission. The item writer has access to the test specifications on a wiki site, which also includes example items and templates for new items. Item writers submit their items via email and receive an acknowledgement that the items have been received.

4.1.2.2 The quality review process

The submitted items are reviewed against a set of checklists derived from the specifications. Items are annotated by two independent reviewers, using a number code system. This identifies any element of the item that does not meet any part of the specifications. Items that pass the quality review stage are added to the computer-based authoring system used for the creation and storage of all Aptis test tasks. Items that do not pass the quality review are returned to item writers with the annotations. In some circumstances, item writers might be asked to revise such items and resubmit, but this is not done as standard practice. In cases where items fail to meet the specifications in only minor detail, the item will be accepted and the necessary changes will be made by the production team. Item writers are informed which of their items have passed the quality review process and have been accepted for further use. All items from receptive skills components are subject to pre-testing before final availability for use in live tests, and item writers do not have knowledge of which items proceed from pre-testing to live test construction, or if any of their items are eventually used in live tests.

4.1.2.3 The pre-testing process

Tasks and items for pre-testing are authored in the CBT authoring system that acts as a repository for all Aptis tasks and items. They are given a workflow status within this system which denotes that they are ready for pre-testing. Audio for the listening and speaking components is recorded in the UK under the supervision of a member of the Aptis team to ensure that appropriate speech rate and timings are adhered to. Tasks are published from the authoring system to the test creation system, and become available there for incorporation into the tests. Sets of tasks and sets of items for pre-testing are constructed using the CBT test creation system. These test versions are reviewed in the CBT delivery format before being made available for centres participating in pre-testing to schedule. Once the pre-testing period is complete, the data analysis of the items is carried out (see Section 3.3.3.1 for details). A number of pre-set statistical criteria are used to investigate task and item performance. Tasks and items that have met the statistical performance criteria are selected for use in operational versions of the test, and these are given a workflow status of 'live' in the authoring system.

4.1.2.4 The production of new versions for use in live administrations

Live versions are created in the integrated CBT delivery system and reviewed in the CBT delivery format before being made available for participating centres to schedule as live tests. The new versions, as noted above, are constructed according to the test specifications for each component, which denote the number of tasks and items at pre-determined levels of difficulty, the total time, etc. All versions are constructed to be comparable in terms of empirical difficulty. As noted in Section 3.3.2.1, pre-testing of the receptive skills components utilises Rasch equating procedures to place all items for a particular component on a common scale for that component. Items selected for use in live test versions thus have known statistical properties, including Rasch logit estimates on a common scale of difficulty. The overall difficulty of test versions can thus be controlled at the version construction stage to ensure that the scores reported to candidates are comparable across versions.

4.1.2.5 Item writer recruitment and training

As noted above, only trained item writers are offered commissions to submit items for the test production cycle. All item writers are trained according to standardised procedures to ensure they are familiar with guidelines for good practice in the fields of testing and item writing, and with the specifications of the Aptis test system.

The original model for ensuring a sufficient pool of trained item writers recruited potential item writers from British Council staff who had completed the Certificate in the Theory and Practice of Language Testing from the University of Roehampton, a distance course of 100 hours over six months. Participants primarily came from teaching centres and exam centres. Participants on that course were invited to put themselves forward for item writer training. Those who accepted were given five days (35 hours) of face-to-face training on all test components (Core, Listening, Reading, Writing, and Speaking). The training involved instruction and hands-on item writing with a combination of peer and instructor review. Following the training, item writers produced example test items during a probationary period. These items were quality reviewed, and item writers were given feedback via email. Item writers who successfully completed the probationary period were invited to become contracted item writers.

New models of item writer training are being introduced in which completion of the Theory and Practice of Language Testing Certificate is not a requirement, provided that participants can demonstrate sufficient experience in language teaching and assessment. One form of training has involved the use of Skype and online file sharing resources to allow training to be delivered by instructors from a distance in conjunction with an instructor present in the room. The various approaches to training item writers make use of the lessons learned from the delivery of training to large numbers of item writers internationally. Lessons learned from the ongoing quality review process in the test production cycle have also been fed back into training, and the insights of item writers have informed the ongoing review and revision of task specifications. Regardless of the mode of delivery of the training, the core elements are standardised to provide item writers with comprehensive training in key concepts in testing important for the process of item writing and review, familiarisation with the CEFR and the test and task specifications for Aptis, as well as providing hands-on practice at item writing and review.

4.2 Accommodations

As described in Section 3.2.1, Aptis General is offered directly to organisations who wish to use it to test their employees, students, etc. Individuals do not register to take the test. As such, organisations are expected to engage in a discussion with the British Council to identify any specific needs of their test-takers which may impact on the ability of the test to derive fair and reliable results. Certain accommodations, if deemed appropriate, can be undertaken from the options already available within the system, while other adjustments are considered on a case-by-case basis.

Accommodations are currently available through the following options:

- different delivery modes for some candidates (e.g., pen and paper over CBT)
- braille versions of the Core and Reading components
- in CBT mode, the colour settings on the screen can be changed for colour settings most appropriate for visually impaired candidates
- extra time can be allocated for candidates in specially prepared CBT versions when this is deemed appropriate.

Other accommodations, such as to the presentation of test content, the format of the response provided by the candidate, or to the testing environment are considered on a case-by-case basis in consultation with the British Council.

4.3 Overview of other documentation on research and validation

Aptis General has been developed within the Aptis test system, a coherent approach to test design, development and production which utilises an explicit model of test development and validation to provide the theoretical framework to drive validation research (see Section 2.2). Aptis General was the first test within the Aptis system to be developed employing this approach. The initial design and development of the test are documented in a series of technical reports which are available online (O'Sullivan, 2015a, 2015b, 2015c – see www.britishcouncil.org/exam/aptis/research/publications).

Validation is an ongoing process, which extends beyond the development stage and continues throughout the live production cycle of a test. An active research agenda is pursued by the British Council to both contribute to the growing body of evidence supporting the uses and interpretations of tests developed within the Aptis test system, and also to inform the revision and ongoing development of the tests to ensure that they reflect the latest research in the field of language testing, and are appropriate for the real-world uses and interpretations to which the tests are put.

The Assessment Research Group at the British Council coordinates validation research. It is carried out through two complementary research strands: the first covers research carried out directly or in collaboration with the Assessment Research Group; the second strand covers research supported through the Assessment Research Awards and Grants (ARAGs) scheme operated by the British Council. The first strand of research is published as a series of Aptis Technical Reports, and the second is published as a series of Research Reports. Both series of reports are made freely available online. For the most recent information regarding proposals which have been accepted under the ARAGs scheme, major research projects being undertaken by the Assessment Research Group, and for completed reports in both the Technical Reports and Research Reports series, readers are referred to the research section of the Aptis website – www.britishcouncil.org/exam/aptis/research

The Assessment Research Group is also engaged in the ongoing analysis and evaluation of operational test data to monitor the statistical performance of live versions of the test. The Assessment Research Group works closely with the Aptis production team to evaluate the statistical performance of live tasks and tests to support the procedures in place for ensuring comparability described in Sections 3.3.2.1, 3.3.3.5 and 4.1.2.

An Assessment Advisory Board, consisting of external experts in language testing and assessment, reviews and evaluates the full program of research and validation coordinated and carried out by the Assessment Research Group. Information on the Board is also available on the Aptis website.

References

American Educational Research Association, American Psychological Association and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, Cambridge University Press.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2, 1–34.

Bachman, L. F., and Palmer, A.S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Chalhoub-Deville, M. (2003). Second language interaction: current perspectives and future trends. *Language Testing* 20(4), 369–383.

Chalhoub-Deville, M. and O'Sullivan, B. (2015). *Validity*. Manuscript in progress.

Chapelle, C. A., Enright, M. K. and Jamieson, J. M. (Eds.) (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York: Routledge.

Chapelle, C. A., Enright, M. K. and Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice,* 29(1), 3–13.

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.

Davidson, F. and Fulcher, G. (2007). The Common European Framework of Reference (CEFR) and the design of language tests: A matter of effect. *Language Teaching*, 40, 231–241. Copenhagen, Denmark.

Dunlea, J. and Fairbairn, J. (2015). *Revising and validating the rating scales for the Aptis Speaking and Writing tests*. Aptis Technical Report. London: British Council. Manuscript in progress.

European Association for Language Testing and Assessment (EALTA). (2006). *Guidelines for Good Practice in Language Testing and Assessment.* Retrieved from: http://www.ealta.eu.org/guidelines.htm

Fairbairn, J. (2015). *Maintaining marking consistency in a large-scale international test: The Aptis experience*. Poster presented at the 12th Annual EALTA Conference.

Field, J. (2015). *Listening*. Manuscript in progress.

Fulcher, G. and Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. New York: Routledge.

Geranpayeh, A. and Taylor, L. (Eds.) (2013). *Examining listening: Research and practice in assessing second language listening*. Cambridge: Cambridge University Press.

Hatch, E. and Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. Boston: Heinle & Heinle.

International Language Testing Association (ILTA). (2007). *Guidelines for practice*. Retrieved from: http://www.iltaonline.com/images/pdfs/ILTA_Guidelines.pdf

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

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38, 319–342.

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, 21, 31–41.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement,* 50, 1–73.

Khalifa, H. and Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge: Cambridge University Press.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement (3rd ed.)*, pp. 13–103. New York: Macmillan.

Milton, J. (2010). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In Bardel, C., Lindqvist, C. and Laufer, B. (Eds), *L2 Vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis.* Eurosla monographs Series, Volume 2. Online: Eurosla.

North, B., Ortega, A. and Sheehan, S. (2010). *A Core Inventory of General English*. British Council / EAQUALS.

O'Sullivan, B. (2000a). *Towards a model of performance in oral language tests*. (Unpublished Ph.D. thesis.) University of Reading.

O'Sullivan, B. (2009). *City and Guilds Communicator IESOL Examination (B2) CEFR linking project*. London: City and Guilds.

O'Sullivan, B. (2011a). Language testing. In J. Simpson (Ed.), *Routledge handbook of applied linguistics*. Oxford: Routledge.

O'Sullivan, B. (2011b). The City and Guilds Communicator examination linking project: a brief overview with reflections on the process. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual*. Cambridge: Cambridge University Press.

O'Sullivan, B. (2015a). *Aptis test development approach.* Aptis Technical Report, TR/2015/001. London: British Council.

O'Sullivan, B. (2015b). *Linking the Aptis reporting scales to the CEFR*. Aptis Technical Report, TR/2015/003. London: British Council.

O'Sullivan, B. (2015c). *Aptis formal trials feedback reports*. Aptis Technical Report, TR/2015/002. London: British Council.

O'Sullivan, B. and Chalhoub-Deville, M. (2015). *Localisation*. Manuscript in progress.

O'Sullivan, B. and Weir, C. J. (2011). Language testing and validation. In B. O'Sullivan (Ed.) *Language testing: theory & practice* (pp.13–32). Oxford: Palgrave.

O'Sullivan, B., Weir, C. & Saville, N. 2002. Using observation checklists to validate speaking-test tasks. *Language Testing,* 19 (1): 33-56.

Shaw, S. and Weir, C J. (2007). *Examining writing: Research and practice in assessing second language writing.* Cambridge: Cambridge University Press and Cambridge ESOL.

Shiotsu, T. (2010). *Components of L2 reading.* Cambridge*:* Cambridge University Press and Cambridge ESOL.

Taylor, L. (Ed.) (2012). *Examining speaking: Research and practice in assessing second language speaking.* Cambridge: Cambridge University Press.

van Zeeland, H. and Schmitt, N. (2012). Lexical coverage and L1 and L2 listening comprehension: the same or different from reading comprehension? *Applied Linguistics,* 2012: 1–24.

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

Weir, C. J. and Milanovic, M. (Eds.) (2003). *Continuity and innovation: a history of the CPE Examination 1913–2002*. Cambridge: Cambridge University Press.

Wu, R. Y. F. (2014). *Validating second language reading examinations: Establishing the validity of the GEPT through alignment with the Common European Framework of Reference*. Cambridge: Cambridge University Press.

Appendix A: Global scale CEFR

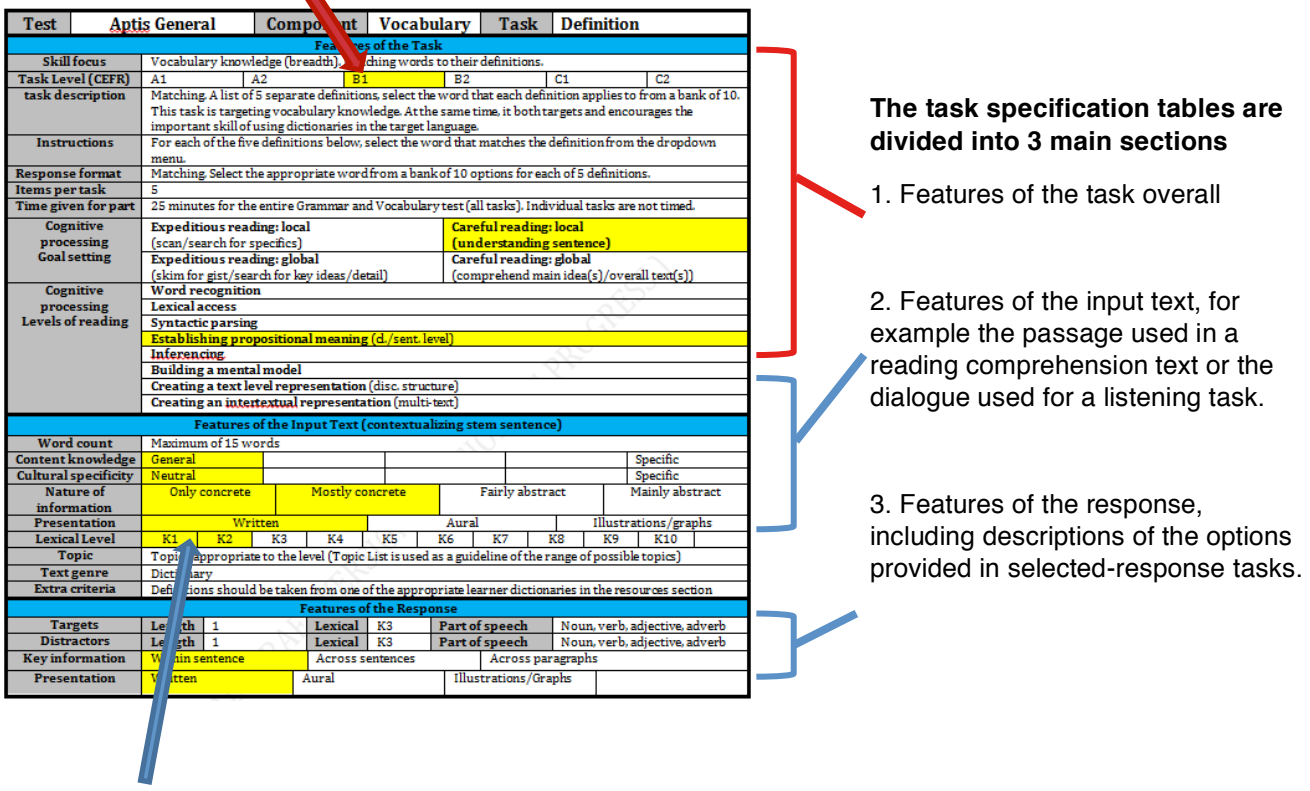
Proficient User	**C2**	Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.
	C1	Can understand a wide range of demanding, longer texts and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.
Independent User	**B2**	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
	B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics, which are familiar, or of personal interest. Can describe experiences and events, dreams, hopes and ambitions, and briefly give reasons and explanations for opinions and plans.
Basic User	**A2**	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.
	A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others, and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

How to read the task specifications tables in the following appendices

The specifications have been designed to incorporate features relevant for describing test tasks proposed in O'Sullivan (2015a), O'Sullivan and Weir (2011) and Weir (2005). The task specifications include both contextual and cognitive parameters for describing tasks. More information on many of these features, and in particular on the models of cognitive processing for the different skills which have been incorporated into these specifications, can be found in Geranpayeh and Taylor (2013), Khalifa and Weir (2007), Shaw and Weir (2009), and Taylor (2012).

Aspects highlighted in yellow

Some categories have a fixed number of alternatives, e.g. the CEFR level targeted by a task. The relevant alternative is highlighted in yellow. In this case, the CEFR level of the task is B1.



Test	Aptis General	Component	Vocabulary	Task	Definition
Features of the Task					
Skill focus	Vocabulary knowledge (breadth), matching words to their definitions.				
Task Level (CEFR)	A1	A2	B1	B2	C1	C2
task description	Matching. A list of 5 separate definitions, select the word that each definition applies to from a bank of 10. This task is targeting vocabulary knowledge. At the same time, it both targets and encourages the important skill of using dictionaries in the target language.				
Instructions	For each of the five definitions below, select the word that matches the definition from the dropdown menu.				
Response format	Matching. Select the appropriate word from a bank of 10 options for each of 5 definitions.				
Items per task	5				
Time given for part	25 minutes for the entire Grammar and Vocabulary test (all tasks). Individual tasks are not timed.				
Cognitive processing Goal setting	**Expeditious reading: local** (scan/search for specifics)	**Careful reading: local (understanding sentence)**			
	Expeditious reading: global (skim for gist/search for key ideas/detail)	**Careful reading: global** (comprehend main idea(s)/overall text(s))			
Cognitive processing Levels of reading	**Word recognition**				
	Lexical access				
	Syntactic parsing				
	Establishing propositional meaning (cl./sent. level)				
	Inferencing				
	Building a mental model				
	Creating a text level representation (disc. structure)				
	Creating an intertextual representation (multi-text)				
Features of the Input Text (contextualizing stem sentence)					
Word count	Maximum of 15 words				
Content knowledge	General				Specific
Cultural specificity	Neutral				Specific
Nature of information	Only concrete	Mostly concrete	Fairly abstract		Mainly abstract
Presentation	Written		Aural		Illustrations/graphs
Lexical Level	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
Topic	Topic appropriate to the level (Topic List is used as a guideline of the range of possible topics)				
Text genre	Dictionary				
Extra criteria	Definitions should be taken from one of the appropriate learner dictionaries in the resources section				
Features of the Response					
Targets	Length	1	Lexical	K3	Part of speech	Noun, verb, adjective, adverb
Distractors	Length	1	Lexical	K3	Part of speech	Noun, verb, adjective, adverb
Key information	Within sentence	Across sentences	Across paragraphs		
Presentation	Written	Aural	Illustrations/Graphs		

The task specification tables are divided into 3 main sections

1. Features of the task overall

2. Features of the input text, for example the passage used in a reading comprehension text or the dialogue used for a listening task.

3. Features of the response, including descriptions of the options provided in selected-response tasks.

Lexical levels

The lexical levels of the input texts and expected response etc., are specified using the BNC-20 lists derived from the British National Corpus by Paul Nation (2006) and adapted by Tom Cobb (http://www.lextutor.ca/freq/eng/). The lists comprise 20 levels, each with 1,000 word families. K1 refers to the most frequent 1,000 word families, K2, the next most frequent 1,000 word families, etc.

List of task specification tables in the following appendices

Appendix B: Task specifications for Aptis General Core component

1. Multiple choice sentence completion
2. Synonym
3. Meaning in context
4. Definition
5. Collocation

Appendix C: Task specifications for Aptis General Reading component

1. Multiple choice gap-fill
2. Sentence re-ordering
3. Bank-filled gap
4. Matching headings to text

Appendix D: Task specifications for Aptis General Listening component

1. MCQ A1
2. MCQ A2
3. MCQ B1
4. MCQ B2

Appendix E: Task specifications for Aptis General Speaking component

1. Speaking Task 1
2. Speaking Task 2
3. Speaking Task 3
4. Speaking Task 4

Appendix F: Task specifications for Aptis General Writing component

1. Writing Task 1
2. Writing Task 2
3. Writing Task 3
4. Writing Task 4

Appendix B: Task specifications for Aptis General Core component

Task: Multiple choice sentence completion

Test	Aptis General	Component	Grammar	Task	Multiple choice sentence completion

Features of the Task						
Skill focus	**Syntax and word usage**					
Task level (CEFR)	A1	A2	B1	B2	C1	C2
Task description	Sentence completion. Select the best word(s) to complete a sentence based on syntactic appropriacy.					
Response format	3-option multiple choice					
Items per task	1 (there is only one gap to fill in each task, making *task* and *item* functionally equivalent for Grammar)					
Time given for part	25 minutes for the entire grammar and vocabulary test. Individual tasks are not timed.					
Cognitive processing Goal setting	**Expeditious reading: local** (scan/search for specifics)			**Careful reading: local** (understanding sentence)		
	Expeditious reading: global (skim for gist/search for key ideas/detail)			**Careful reading: global** (comprehend main idea(s)/overall text(s))		
Cognitive processing Levels of reading	**Word recognition**					
	Lexical access					
	Syntactic parsing					
	Establishing propositional meaning (cl./sent. level)					
	Inferencing					
	Building a mental model					
	Creating a text level representation (disc. structure)					
	Creating an intertextual representation (multi-text)					

Features of the Input Text											
Word count	A1 items maximum of 8 words. A2-B2 items maximum of 15 words.										
Content knowledge (A1-B2)	General										Specific
Cultural specificity (A1-B2)	Neutral										Specific
Nature of information A1	Only concrete	Mostly concrete			Fairly abstract			Mainly abstract			
Nature of information A2	Only concrete	Mostly concrete			Fairly abstract			Mainly abstract			
Nature of information B1	Only concrete	Mostly concrete			Fairly abstract			Mainly abstract			
Nature of information B2	Only concrete	Mostly concrete			Fairly abstract			Mainly abstract			
Presentation	Verbal			Non-verbal (i.e. graphs)				Both			
Lexical level A1 target	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10	
Lexical level A2 target	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10	
Lexical Level B1 target	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10	
Lexical level B2 target	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10	
Topic	Topics appropriate to the level. (Topic List is used as a guideline of the range of possible topics.)										
Genre	As stand-alone sentences, it is difficult to identify a specific genre. However, the sentences should be plausible extracts from the range of texts likely to be encountered by candidates in the TLU domain for Aptis General. Some elements of spoken grammar will be targeted with dialogues.										

Features of the Response					
Target	**Length**	1–3 words	**Lexical**	Same as the level for the stem sentence	
Target (grammatical level)	Targets will be chosen from grammatical exponents for the targeted level from the British Council Equals Core Inventory.				
Distractors	**Length**	1–3 words	**Lexical**	Same as the level for the stem sentence	
Key information	Within sentence	Across sentences	Across paragraphs		
Extra criteria	All of the options must be plausible as stand-alone words outside the stem. It should not be possible to rule out an option without reference to the stem based on spelling or non-existent morphology				
Presentation	Written	Aural	Illustrations/Graphs		

Task: Synonym

Test	Aptis General	Component	Vocabulary	Task	Synonym

Features of the Task						
Skill focus	**Vocabulary knowledge (breadth). Matching words with the same or similar meanings.**					
Task level (CEFR)	A1	A2	B1	B2	C1	C2
Task description	Word matching. Match two words which have the same or very similar meanings. For each of 5 target words, select the best match from a bank of 10 options					
Instructions to candidates	Select a word from the list that has the same or a very similar meaning to the word on the left. *(This is slightly different to present rubric)*					
Response format	Matching from a bank of options. For 5 target words, select the best match for each from a bank of 10 options					
Items per task	5					
Time given for part	25 minutes for the entire Grammar and Vocabulary test (all tasks). Individual tasks are not timed.					
Cognitive processing Levels of reading	**Word recognition**					
	Lexical access					
	Syntactic parsing					
	Establishing propositional meaning (cl./sent. level)					
	Inferencing					
	Building a mental model					
	Creating a text level representation (disc. structure)					
	Creating an intertextual representation (multi-text)					

Features of the Response						
Target	**Length**	1	**Lexical**	K1	**Part of speech**	Nouns, verbs, adjectives
Distractors	**Length**	1	**Lexical**	K1	**Part of speech**	Nouns, verbs, adjectives
Presentation	Written		Aural		Illustrations/Graphs	

Task: Meaning in context

Test	Aptis General	Component	Vocabulary	Task	Meaning in context

Features of the Task						
Skill focus	**Vocabulary knowledge (breadth). Understanding meaning from context**					
Task level (CEFR)	A1	A2	B1	B2	C1	C2
Task description	Sentence completion. For 5 stand-alone sentences (i.e. the sentences do not form a text), select the best option form a bank of 10 to complete each sentence. The correct word will be the most appropriate and plausible lexical choice for the context.					
Instructions	Complete each sentence using a word from the dropdown list.					
Response format	Matching. Select the best option for each target sentence from a bank of 10.					
Items per task	5					
Time given for part	25 minutes for the entire Grammar and Vocabulary test (all tasks). Individual tasks are not timed.					
Cognitive processing Goal setting	**Expeditious reading: local** (scan/search for specifics)			**Careful reading: local** (understanding sentence)		
	Expeditious reading: global (skim for gist/search for key ideas/detail)			**Careful reading: global** (comprehend main idea(s)/overall text(s))		
Cognitive processing Levels of reading	**Word recognition**					
	Lexical access					
	Syntactic parsing					
	Establishing propositional meaning (cl./sent. level)					
	Inferencing					
	Building a mental model					
	Creating a text level representation (disc. structure)					
	Creating an intertextual representation (multi-text)					

Features of the Input Text											
Word count	Maximum 15										
Content knowledge	General										Specific
Cultural specificity	Neutral										Specific
Nature of information	Only concrete			Mostly concrete			Fairly abstract			Mainly abstract	
Presentation	Written					Aural				Illustrations/graphs	
Lexical level A2	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10	
Lexical level B1	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10	
Topic	Topics appropriate to the level. (Topic List is used as a guideline of the range of possible topics.)										
Text genre	As stand-alone sentences, it is difficult to identify a specific genre. However, the sentences should be plausible extracts from the range of texts likely to be encountered by candidates in the TLU domain for Aptis General, and relevant to the level.										

Features of the Response						
Target A2	**Length**	1	**Lexical**	K2	**Part of speech**	Nouns, verbs, adjectives
Distractors A2	**Length**	1	**Lexical**	K2	**Part of speech**	Nouns, verbs, adjectives
Target B1	**Length**	1	**Lexical**	K3	**Part of speech**	Nouns, verbs, adjectives
Distractors B1	**Length**	1	**Lexical**	K3	**Part of speech**	Nouns, verbs, adjectives
Key information	Within sentence		Across sentences		Across paragraphs	
Presentation	Written		Aural		Illustrations/Graphs	

Task: Definition

Test	Aptis General	Component	Vocabulary	Task	Definition

Features of the Task						
Skill focus	**Vocabulary knowledge (breadth). Matching words to their definitions.**					
Task level (CEFR)	A1	A2	B1	B2	C1	C2
Task description	Matching. A list of 5 separate definitions, select the word that each definition applies to from a bank of 10. This task is targeting vocabulary knowledge. At the same time, it both targets and encourages the important skill of using dictionaries in the target language.					
Instructions	For each of the 5 definitions below, select the word that matches the definition from the dropdown menu.					
Response format	Matching. Select the appropriate word from a bank of 10 options for each of 5 definitions.					
Items per task	5					
Time given for part	25 minutes for the entire Grammar and Vocabulary test (all tasks). Individual tasks are not timed.					
Cognitive processing Goal setting	**Expeditious reading: local** (scan/search for specifics)			**Careful reading: local** (understanding sentence)		
	Expeditious reading: global (skim for gist/search for key ideas/detail)			**Careful reading: global** (comprehend main idea(s)/overall text(s))		
Cognitive processing Levels of reading	**Word recognition**					
	Lexical access					
	Syntactic parsing					
	Establishing propositional meaning (cl./sent. level)					
	Inferencing					
	Building a mental model					
	Creating a text level representation (disc. structure)					
	Creating an intertextual representation (multi-text)					

Features of the Input Text (contextualising stem sentence)											
Word count	Maximum of 15 words										
Content knowledge	General				Specific						
Cultural specificity	Neutral				Specific						
Nature of information	Only concrete	Mostly concrete	Fairly abstract	Mainly abstract							
Presentation	Written	Aural	Illustrations/graphs								
Lexical level	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10	
Topic	Topics appropriate to the level. (Topic List is used as a guideline of the range of possible topics.)										
Text genre	Dictionary										
Extra criteria	Definitions should be taken from one of the appropriate learner dictionaries in the resources section.										

Features of the Response						
Targets	**Length**	1	**Lexical**	K3	**Part of speech**	Noun, verb, adjective, adverb
Distractors	**Length**	1	**Lexical**	K3	**Part of speech**	Noun, verb, adjective, adverb
Key information	Within sentence	Across sentences	Across paragraphs			
Presentation	Written	Aural	Illustrations/Graphs			

Task: Collocation

Test	Aptis General	Component	Vocabulary	Task	Collocation

Features of the Task						
Skill focus	**Vocabulary knowledge (depth). For words targeted from the appropriate vocabulary level, understanding how those lexical items operate in context and what other lexical items will likely be used with them.**					
Task level (CEFR)	A1	A2	B1	B2	C1	C2
Task description	Word matching. For a list of 5 target words, select the word which is most commonly used with the target word from a list of 10 options. The collocation pairs would be used in a direct sequence. This task targets depth of vocabulary knowledge regarding the word targeted. It is not simply knowledge of the general meaning or semantic field, but in-depth knowledge about how the word is used in context.					
Instructions	Select a word from the list that is most often used with the word on the left.					
Response format	Matching. For each of 5 target words, select the best option from a bank of 10.					
Items per task	5					
Time given for part	25 minutes for the entire reading test (all tasks). Individual tasks are not timed.					
Cognitive processing Levels of reading	**Word recognition**					
	Lexical access					
	Syntactic parsing					
	Establishing propositional meaning (cl./sent. level)					
	Inferencing					
	Building a mental model					
	Creating a text level representation (disc. structure)					
	Creating an intertextual representation (multi-text)					

Features of the Response						
Target	**Length**	1	**Lexical**	K4-K5	**Part of speech**	Nouns, verbs, adjectives, adverbs
Distractors	**Length**	1	**Lexical**	K1-K4	**Part of speech**	Nouns, verbs, adjectives, adverbs
Presentation	Written		Aural		Illustrations/Graphs	

Appendix C: Task specifications for Aptis General Reading component

Task: Multiple choice gap-fill

Test	**Aptis General**	**Component**	**Reading**	**Task**	**Multiple choice gap-fill**

Features of the Task						
Skill focus	**Reading comprehension up to the sentence level**					
Task level (CEFR)	A1	A2	B1	B2	C1	C2
Task description	Multiple-choice gap fill. A short text of 6 sentences is presented. Each sentence contains one gap. Test-takers choose the best option from a pull-down menu for each gap to complete the sentence. The first sentence is an example with the gap completed.					
Instructions to candidates	*(The text in brackets will vary according to the specific content of the task.)* Read the (*letter, email, postcard, note, memo*) from (*writer's relationship to reader*). Choose one word from the list for each gap. The first one is done from you.					
Response format	3-option multiple choice					
Items per task	5					
Time given for part	30 minutes for the entire reading test (all tasks). Individual tasks are not timed.					
Cognitive processing Goal setting	**Expeditious reading: local** (scan/search for specifics)			**Careful reading: local** (understanding sentence)		
	Expeditious reading: global (skim for gist/search for key ideas/detail)			**Careful reading: global** (comprehend main idea(s)/overall text(s))		
Cognitive processing Levels of reading	**Word recognition**					
	Lexical access					
	Syntactic parsing					
	Establishing propositional meaning (cl./sent. level)					
	Inferencing					
	Building a mental model					
	Creating a text level representation (disc. structure)					
	Creating an intertextual representation (multi-text)					

Features of the Input Text										
Word count	50-60 words				**Number of sentences (total)**		6			
Avg sentence length	8-10 (This is an average figure. Individual sentences will span a range above and below the average.)									
Domain	Public		Occupational		Educational		Personal			
Discourse mode	Descriptive	Narrative		Expository		Argumentative		Instructive		
Content knowledge	General						Specific			
Cultural specificity	Neutral						Specific			
Nature of information	Only concrete		Mostly concrete		Fairly abstract		Mainly abstract			
Presentation	Verbal			Non-verbal (i.e. graphs)			Both			
Lexical level	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
Grammatical level	A1 Grammatical exponents (See Guidelines on Adhering to Grammatical Level)									
Topic	From topic list for A1. (For personal notes and letters, no one topic may be dominant, and a number of different topics may be referred to in the process of providing an update on daily events, etc.)									
Text genre	Emails, letters, notes, postcards									
Intended writer/reader relationship	The writer is known to the intended reader, and will be part of the typical network of family and friends relevant to the A1 field of activity. The relationship is specified in the rubric.									

Features of the Response						
Target	**Length**	1 word	**Lexical**	K1	**Part of speech**	Noun, verb, adjective
Distractors	**Length**	1 word	**Lexical**	K1	**Part of speech**	Noun, verb, adjective
Key information	Within sentence		Across sentences		Across paragraphs	
Presentation	Written		Aural		Illustrations/graphs	

Task: Sentence re-ordering

Test	Aptis General	Component	Reading	Task	Sentence re-ordering

Features of the Task						
Skill focus	**Inter-sentence cohesion**					
Task level (CEFR)	A1	A2	B1	B2	C1	C2
Task description	Re-order jumbled sentences to form a short, cohesive text. Seven sentences are presented, with the introductory sentence given first in the right order. The remaining sentences must be re-ordered to form a short text which tells a story or describes something as a simple list of points or actions which would hang together as a text in a linear sequence.					
Instructions to candidates	*(The text in brackets will vary according to the specific content of the task.)* The sentences below are from a (*newspaper story, instructions for a task, directions*). Put the sentences in the right order. The first sentence is done for you.					
Response format	Re-ordering of fixed number (6) of jumbled sentences.					
Items per task	6 (each sentences is counted as a single item)					
Time given for part	30 minutes for the entire reading test (all tasks). Individual tasks are not timed.					
Cognitive processing Goal setting	**Expeditious reading: local** (scan/search for specifics)			**Careful reading: local** (understanding sentence)		
	Expeditious reading: global (skim for gist/search for key ideas/detail)			**Careful reading: global** (comprehend main idea(s)/overall text(s))		
Cognitive processing Levels of reading	**Word recognition**					
	Lexical access					
	Syntactic parsing					
	Establishing propositional meaning (cl./sent. level)					
	Inferencing					
	Building a mental model					
	Creating a text level representation (disc. structure)					
	Creating an intertextual representation (multi-text)					

Features of the Input Text										
Word count	90–100 words	**Average sentence length**	7 (1 introductory sentence + 6 jumbled sentences)							
Avg sentence length	13–15 (This is an average figure calculated across the whole text. Individual sentences will span a range above and below the average.)									
Domain	Public	Occupational	Educational	Personal						
Discourse mode	Descriptive	Narrative	Expository	Argumentative	Instructive					
Content knowledge	General				Specific					
Cultural specificity	Neutral				Specific					
Nature of information	Only concrete	Mostly concrete	Fairly abstract	Mainly abstract						
Presentation	Written	Aural	Illustrations/graphs							
Lexical level	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
Lexical level	All vocabulary should be from within the K1 and K2 levels.									
Readability	Flesch Kincaid of 4–6 (approximate guidelines only, as readability estimates generally require texts of 200 words or more for stable estimates)									
Topic	From topic list for A2									
Text genre	Newspapers, notices and regulations, instruction manuals, instructional materials (e.g. homework or assignment instructions, textbook extracts describing historical events or biographies). The texts are adapted to the level. Although not intended to be authentic, they should reflect features of relevant texts from the TLU domain. It should be possible to answer the questions: *Where would a reader be likely to see a text like this outside the test? Is the genre relevant to TLU tasks important for Aptis General test-takers at A2 level?*									
Intended writer/reader relationship	The relationship is not specified. Many texts (e.g. newspaper articles, instructions) will be written for a general audience and not a specific reader.									

Features of the Response				
Target	**Length**	Sentence length (as per features of the text above)	**Lexical**	As per text above
Key information	Within sentence	Across sentences	Across paragraphs	
Presentation	Written	Aural	Illustrations/graphs	

Task: Bank-filled gap

Test	Aptis General	Component	Reading	Task	Banked gap-fill

Features of the Task						
Skill focus	**Text level reading comprehension, integrating propositions across a short text into a discourse-level representation.**					
Task level (CEFR)	A1	A2	B1	B2	C1	C2
Task description	Banked gap-fill. Candidates read a short expository text and choose the most appropriate word from a bank of options to fill seven gaps in the text. The bank of options includes the 7 targeted words and 3 distractors.					
Instructions to candidates	Read the text and complete each gap with a word from the list at the bottom of the page.					
Response format	Banked gap-fill. Seven target words are selected from a bank of 10 options					
Items per task	7					
Time given for part	30 minutes for the entire reading test (all tasks). Individual tasks are not timed.					
Cognitive processing Goal setting	**Expeditious reading: local** (scan/search for specifics)			**Careful reading: local** (understanding sentence)		
	Expeditious reading: global (skim for gist/search for key ideas/detail)			**Careful reading: global** (comprehend main idea(s)/overall text(s))		
Cognitive processing Levels of reading	**Word recognition**					
	Lexical access					
	Syntactic parsing					
	Establishing propositional meaning (cl./sent. level)					
	Inferencing					
	Building a mental model					
	Creating a text level representation (disc. structure)					
	Creating an intertextual representation (multi-text)					

Features of the Input Text										
Word count	140–160 words (including target words for gaps)				**Number of sentences**	Not specified				
Avg sentence length	13–15 (This is an average figure. Individual sentences will span a range above and below the average.)									
Domain	Public		Occupational		Educational		Personal			
Discourse mode	Descriptive		Narrative		Expository		Argumentative		Instructive	
Content knowledge	General								Specific	
Cultural specificity	Neutral								Specific	
Nature of information	Only concrete		Mostly concrete		Fairly abstract		Mainly abstract			
Presentation	Verbal				Non-verbal (i.e. graphs)				Both	
Lexical level	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
Lexical level	The cumulative coverage should reach 95% at the K3 level. No more than 5% of words should be beyond K3.									
Readability	Flesch Kincaid grade level of 6–8 (approximate guidelines only, as readability estimates generally require texts of 200 words or more for stable estimates)									
Topic	From topic list for B1.									
Text genre	Magazines, newspapers, instructional materials (such as extracts from textbooks describing important events or people). Although short biographies lend themselves well to this task, it is important to have a range of texts describing events, locations, concrete processes or activities, etc., in addition to biographical descriptions. The texts are adapted to the level. Although not intended to be authentic, they should reflect features of relevant texts from the TLU domain. It should be possible to answer the questions: *Where would a reader be likely to see a text like this outside the test? Is the genre relevant to TLU tasks important for Aptis General test-takers at B1 level?*									
Writer/reader relationship	The relationship is not specified. The texts will typically be written for a general audience, not a specific reader.									

Features of the Response						
Target	**Length**	1 word	**Lexical**	K1-K3	**Part of Speech**	Noun, verb, adjective
Distractors	**Length**	1 word	**Lexical**	K1-K3	**Part of Speech**	Noun, verb, adjective
Key information	Within sentence		Across sentences		Across paragraphs	
Presentation	Written		Aural		Illustrations/graphs	

Task: Matching headings to text

Test	Aptis General	Component	Reading	Task	Matching headings to text

Features of the Task						
Skill focus	**Expeditious global reading of a longer text, integrating propositions across a longer text into a discourse-level representation.**					
Task level (CEFR)	A1	A2	B1	B2	C1	C2
Task description	Matching headings to paragraphs within a longer text. Candidates read through a longer text consisting of 7 paragraphs, identifying the best heading for each paragraph from a bank of 8 options.					
Instructions to candidates	Read the passage quickly. Choose the best heading for each numbered paragraph (1-7) from the dropdown box. There is one more heading than you need.					
Response format	Matching headings to paragraphs in a longer text. Select 7 headings from 8 options.					
Items per task	7 (each heading is one item)					
Time given for part	30 minutes for the entire reading test (all tasks). Individual tasks are not timed.					
Cognitive processing Goal setting	**Expeditious reading: local** (scan/search for specifics)			**Careful reading: local** (understanding sentence)		
	Expeditious reading: global (skim for gist/search for key ideas/detail)			**Careful reading: global** (comprehend main idea(s)/overall text(s))		
Cognitive processing Levels of reading	**Word recognition**					
	Lexical access					
	Syntactic parsing					
	Establishing propositional meaning (cl./sent. level)					
	Inferencing					
	Building a mental model					
	Creating a text level representation (disc. structure)					
	Creating an intertextual representation (multi-text)					

Features of the Input Text										
Word count	700–750 words			**Number of sentences**	Not specified					
Avg sentence length	18–20 (This is an average figure. Individual sentences will span a range above and below the average.)									
Domain	Public		Occupational		Educational		Personal			
Discourse mode	Descriptive		Narrative		Expository	Argumentative	Instructive			
Content knowledge	General						Specific			
Cultural specificity	Neutral						Specific			
Nature of information	Only concrete		Mostly concrete		Fairly abstract		Mainly abstract			
Presentation	Verbal			Non-verbal (i.e. graphs)			Both			
Lexical level	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
Lexical level	The cumulative coverage should reach 95% at the K5 level. No more than 5% of words should be beyond the K5 level. (See Guidelines on Adhering to Lexical Level for more information).									
Grammatical level	A1-B2 Grammatical exponents (See Guidelines on Adhering to Grammatical Level)									
Readability	Flesch Kincaid Grade Level of 9–12									
Topic	From topic list for B2.									
Text genre	Magazines, newspapers, instructional materials (such as extracts from undergraduate textbooks describing important events, the ideas, or movements). It should be possible to answer the questions: *Where would a reader be likely to see a text like this outside the test? Is the genre relevant to TLU tasks important for Aptis General test-takers at B2 level?*									
Intended writer/reader relationship	The relationship is not specified. The texts will typically be written for a general audience, not a specific reader.									

Features of the Response						
Targets	**Length**	Up to 10 words	**Lexical**	K1-K5	**Grammatical**	A1-B2
Distractors	**Length**	Up to 10 words	**Lexical**	K1-K5	**Grammatical**	B1-B2
Key information	Within sentence		Across sentences		Across paragraphs	
Presentation	Written		Aural		Illustrations/graphs	

Appendix D: Task specifications for Aptis General Listening component

Task: MCQ A1

Test	Aptis General	Component	Listening	Task	MCQ A1

Features of the Task						
Skill focus	Lexical recognition					
Task level (CEFR)	A1	A2	B1	B2	C1	C2
Task description	Listen to a short monologue and choose the best option to answer a question or complete a statement. The task focuses on identification of a specific word or number in a short message from familiar, everyday life situations.					
Instructions to candidates	The rubric will always contain two parts: 1) a short contextualisation: *listen to the message for Mary from Arturo;* 2) A short question to focus listening: e.g. *What is Arturo's phone number?*					
Presentation	Written	Aural	Illustrations / graphs			
Response format	4-option multiple choice	**Items per task**	1			
Time given for part	50 minutes for the entire Listening test (all tasks). Individual tasks are not timed.					
Kind of information targeted	**Lexical recognition**	**Factual information**				
	Interpretative meaning at the utterance level	**Meaning at discourse level**				
Cognitive processing Levels of listening	**Input decoding**					
	Lexical search					
	Syntactic parsing					
	Meaning construction					
	Discourse construction					

Features of the Input Text										
Length	30 seconds	**Words**	60–80							
Accent	Standard British English speaker likely to be encountered in the UK. Native speakers of English.									
Domain	Public	Occupational	Educational	Personal						
Discourse mode	Descriptive	Narrative	Expository	Argumentative	Instructive					
Pattern	Monologue	Dialogue								
Content knowledge	General				Specific					
Cultural specificity	Neutral				Specific					
Nature of information	Only concrete	Mostly concrete	Fairly abstract	Mainly abstract						
Presentation	Written	Aural	Illustrations/graphs							
Lexical level	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
Lexical level	All vocabulary should be from within the K1 level (See Guidelines on Adhering to Lexical Level)									
Grammatical level	A1 Grammatical exponents (See Guidelines on Adhering to Grammatical Level)									
Topic	See topic list.									
Text genre	Recorded telephone messages: The message may come from situations likely to occur in one of several domains (see above). The speaker will be known to the intended listener, and the information will be limited to concrete, everyday familiar topics.									
Relationship of participants	The speaker will be known to the intended listener, with the specific relationship depending on the domain and genre (e.g. educational: teacher-student; occupational: colleagues; personal: friends or family)									

Features of the Response						
Stem	**Length**	8 (max) words	**Lexical**	K1	**Grammar**	A1 exponents
Presentation	Written	Aural			Illustrations/graphs	
Options	**Length**	1-3 words	**Lexical**	K1	**Grammar**	A1 exponents
Presentation	Written	Aural			Illustrations/graphs	
Key information	Within sentence	Across sentences	Across paragraphs			

Task: MCQ A2

Test	Aptis General	Component	Listening	Task	MCQ A2

Features of the Task						
Skill focus	**Identifying specific, factual information**					
Task level (CEFR)	A1	A2	B1	B2	C1	C2
Task description	Q&A about listening text. Listen to short monologues and conversations to identify short, specific pieces of information.					
Further information						
Instructions to candidates	The rubric will always contain two parts: 1) a short contextualisation*: listen to the message for Mary from Arturo* or *listen to the man and woman talking*; 2) The second part of the rubric will be a short question, e.g. *What is Arturo's phone number*?					
Presentation	Written	Aural	Illustrations/graphs			
Response format	4-option multiple choice	**Items per task**	1			
Time given for part	Approximately 50 minutes for the entire Listening test (all tasks). Individual tasks are not timed.					
Kind of information targeted	**Lexical recognition**	**Factual information**				
	Interpretative meaning at the utterance	**Meaning at discourse level**				
Cognitive processing Levels of listening	**Input decoding**					
	Lexical search					
	Syntactic parsing					
	Meaning construction					
	Discourse construction					

Features of the Input Text											
Length	30 seconds	**Words**	60–80	**speed**	2.2 – 2.6 words per second (approximate)						
Accent	Standard British English speaker likely to be encountered in the UK. Native speakers of English.										
Domain	Public	Occupational	Educational	Personal							
Discourse mode	Descriptive	Narrative	Expository	Argumentative	Instructive						
Pattern	Monologue	Dialogue									
Content knowledge	General				Specific						
Cultural specificity	Neutral				Specific						
Nature of information	Only concrete	Mostly concrete	Fairly abstract	Mainly abstract							
Presentation	Written	Aural	Illustrations / graphs								
Lexical Level	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10	
Lexical Level	All vocabulary should be from within the K1/K2 level (See Guidelines on Adhering to Lexical Level)										
Grammatical level	A2 Grammatical exponents (See Guidelines on Adhering to Grammatical Level)										
Topic	From topic list for A2										
Text genre	Monologues: Recorded telephone messages, instructions, lectures/presentations, public announcements, weather forecasts, news programs, short speeches, advertising. Dialogues: Interpersonal conversations (includes interaction in educational, occupational, and public domains, e.g. conversation between sales assistant and customer, or conversation between two students about study.										
Relationship of participants	Monologues: The speaker may or may not be known to the intended listener. Dialogues: Participants may be known to each other (friends, colleagues, teacher/student) or unknown (sales assistant/customer, public announcement).										

Features of the Response						
Stem	**Length**	8 (max) words	**Lexical**	K1	**Grammar**	A1 exponents
Presentation	Written	Aural	Illustrations/Graphs			
Options	**Length**	1-5 words	**Lexical**	K1	**Grammar**	A1 exponents
Presentation	Written	Aural	Illustrations/Graphs			
Key information	Within utterance/turn	Across utterances/turn				

Task: MCQ B1

Test	Aptis General	Component	Listening	Task	MCQ B1

Features of the Task						
Skill focus	**Identifying factual information**					
Task level (CEFR)	A1	A2	B1	B2	C1	C2
Task description	Q&A about listening text. Listen to short monologues and conversations to identify factual information.					
Instructions to candidates	The rubric will always contain two parts: 1) a short contextualisation: *Listen to the museum guide*. *Listen to the man and woman planning a meeting*; 2) The second part of the rubric may be either a short question (e.g. *What is special about the painting*?) or a short instruction: *(Find out where the meeting will be held…)*					
Response format	4-option multiple choice		**Items per task**	1		
Time given for part	Approximately 50 minutes for the entire Listening test (all tasks). Individual tasks are not timed.					
Kind of information targeted	**Lexical recognition**			**Factual information**		
	Interpretative meaning at the utterance			**Meaning at discourse level**		
Cognitive processing Levels of listening	**Input decoding**					
	Lexical search					
	Syntactic parsing					
	Meaning construction					
	Discourse construction					

Features of the Input Text										
Length	30 seconds	**Words**	70–90	**speed**	2.4 – 3.0 words per second (approximate)					
Accent	Standard British English speaker likely to be encountered in the UK. Native speakers of English.									
Domain	Public		Occupational		Educational		Personal			
Discourse mode	Descriptive	Narrative		Expository		Argumentative		Instructive		
Pattern	Monologue				Dialogue					
Content knowledge	General						Specific			
Cultural specificity	Neutral						Specific			
Nature of information	Only concrete		Mostly concrete		Fairly abstract		Mainly abstract			
Presentation	Written			Aural		Illustrations / graphs				
Lexical level	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
Lexical level	The cumulative coverage should reach 95% at the K3 level. No more than 5% of words should be beyond K3.									
Topic	From topic list for B1.									
Text genre	Monologues: Recorded telephone messages, instructions, lectures/presentations, public announcements, weather forecasts, news programs, short speeches. Dialogues: Interpersonal conversations (i.e. interaction in educational, occupational, and public domains, e.g. conversation between sales assistant and customer, or conversation between two students about study).									
Relationship of participants	Monologues: The speaker may or may not be known to the intended listener. Dialogues: Participants may be known to each other (friends, colleagues, teacher/student) or unknown (sales assistant/customer, public announcement).									

Features of the Response						
Stem	**Length**	10 (max) words	**Lexical**	K1–K2	**Grammar**	A1–A2 exponents
Presentation	Written		Aural		Illustrations/graphs	
Options	**Length**	1–8 words	**Lexical**	K1–K2	**Grammar**	A1–A2 exponents
Presentation	Written		Aural		Illustrations/graphs	
Key information	Within sentence		Across sentences		Across paragraphs	

Task: MCQ B2

Test	Aptis General	Component	Listening	Task	MCQ B2

Features of the Task						
Skill focus	Discourse construction, meaning representation and inference in abstract texts					
Task level (CEFR)	A1	A2	B1	B2	C1	C2
Task description	Q&A about listening text. Listen to monologues and dialogues. Questions will target understanding of the speaker's attitude, opinion, intention, or other information requiring textual inferencing and the integration of propositions across the input text.					
Instructions to candidates	The rubric will always contain two parts: 1) a short contextualisation: *Listen to the lecturer talking about a book*. *Listen to a teacher and a student talking about an assignment*; 2) the second part may be either a short question (e.g. *What is the reason for the book's success?)* or a short instruction (*Find out what the student decides to do…*)					
Response format	4-option multiple choice			**Items per task**	1	
Time given for part	50 minutes for the entire Listening test (all tasks). Individual tasks are not timed.					
Kind of information targeted	**Lexical recognition**			**Factual information**		
	Interpretative meaning at the utterance			**Meaning at discourse level**		
Cognitive processing Levels of listening	**Input decoding**					
	Lexical search					
	Syntactic parsing					
	Meaning construction					
	Discourse construction					

Features of the Input Text										
Length	30 seconds	**Words**	90–110	**Speed**	3.0– 3.6 words per second (approximate)					
Accent	Standard British English speaker likely to be encountered in the UK. Native speakers of English.									
Domain	Public	Occupational	Educational	Personal						
Discourse mode	Descriptive	Narrative	Expository	Argumentative	Instructive					
Pattern	Monologue	Dialogue								
Content knowledge	General				Specific					
Cultural specificity	Neutral				Specific					
Nature of information	Only concrete	Mostly concrete	Fairly abstract	Mainly abstract						
Presentation	Written	Aural	Illustrations/graphs							
Lexical level	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
Lexical level	The cumulative coverage should reach 95% at the K5 level. No more than 5% of words should be beyond K5.									
Topic	From topic list for B2.									
Text genre	Monologues: Recorded telephone messages, instructions, lectures, public announcements, weather forecasts, news programs, short speeches, short features on broadcast media, reviews on TV and radio. Dialogues: Interviews (both live and on broadcast media), debates and discussions, interpersonal conversations (i.e. interaction in educational, occupational, and public domains e.g. conversation between sales assistant and customer, or conversation between professor and student, etc.)									
Relationship of participants	Monologues: The speaker may or may not be known to the intended listener. Dialogues: Participants may be known to each other (friends, colleagues, teacher/student) or unknown (sales assistant/customer, public announcement etc.).									

Features of the Response						
Stem	**Length**	12 words (max)	**Lexical**	K1–K4	**Grammar**	A1–B1 exponents
Presentation	Written	Aural	Illustrations/Graphs			
Options	**Length**	1–10 words	**Lexical**	K1–K4	**Grammar**	A1–B1 exponents
Presentation	Written	Aural	Illustrations/graphs			
Key information	Within sentence	Across sentences	Across paragraphs			

Appendix E: Task specifications for Aptis General Speaking component

Speaking Task 1

Test	Aptis General	Component	Speaking	Task	Task 1
Features of the Task					
Skill focus	**Providing simple personal information and responding to simple spoken questions on familiar topics**				
Task level (CEFR)	A1 / A2		B2	C1	C2
Task description	Candidate responds to three spoken questions on personal topics. Each question is presented separately, and the candidate records his/her spoken response before the next question is presented. The task is designed to elicit short responses to spoken questions on familiar and concrete topics, and the rubric is phrased in the 1st person to approximate interaction with an interlocutor.				
Instructions to candidates	*Part one. In this part, I'm going to ask you three short questions about yourself and your interests. You will have 30 seconds to reply to each question. Begin speaking when you hear this sound (beep).*				
Presentation of rubric	Aural	Written	Other non-verbal (e.g. photo)		
Response format	Q&A	Short turn	Long turn		
Planning time	None				
Delivery	Face-to-face	Telephone	Computer	Other	
Nature of input	Real time (face-to-face)	Real time (remote)	Pre-recorded input	No aural input	
	Unscripted	Guided	Semi-scripted	Scripted	N/A
Nature of interaction	Interlocutor–Candidate (I–C)	Candidate–Candidate (C–C)			
	Candidate only (C)	Interlocutor–Candidate–Candidate			
Functions targeted	**Informational functions**	**Interactional functions**	**Managing interaction**		
	Providing personal information	Agreeing			
	Explaining opinions/preferences	Disagreeing	Initiating		
	Elaborating	Modifying/commenting	Changing topics		
	Justifying opinions	Asking for opinions	Reciprocating		
	Comparing	Persuading	Deciding		
	Speculating	Asking for information			
	Staging	Conversational repair			
	Describing	Negotiation of meaning			
	Summarising				
	Suggesting				
	Expressing preferences				
Features of the Input / Prompt					
Description	3 short questions on familiar personal topics.				
Length of questions	Maximum of 12 words per sentence				
Lexical level	K1 / K2 / K3 / K4 / K5 / K6 / K7 / K8 / K9 / K10				
Grammatical level	A1 Grammatical exponents (See Guidelines on Adhering to Grammatical Level)				
Content knowledge	General				Specific
Cultural specificity	Neutral				Specific
Nature of information	Only concrete	Mostly concrete	Fairly abstract	Mainly abstract	
Relevant domain	Public	Occupational	Educational	Personal	
Topic	From topic list for A1/A2. Appropriate questions will be about familiar, everyday topics that typical Aptis General test-takers can respond to from direct, personal knowledge and experience. The topics will reflect the kind of questions likely to be asked in interaction in the personal domain.				
Features of the Expected Response					
Description	Short responses to 3 questions at the sentence / clause level. Candidate must provide sufficient content in response to at least 2 questions to achieve a rating of 3 (out of 5) for the task.				
Length of response	Up to 30 seconds per question. Adequate responses will extend beyond word/phrase level.				
Lexis/grammar	Demonstration of grammatical control at the A2 level (producing utterances at the clause/sentence level) necessary for a rating of 3 (out of 5) for the task. A1/A2 lexis sufficient to respond adequately to all questions.				
Rating scale for task	A task-specific holistic rating scale is used for the task. The rating scale is a 6-point scale from 0–5. An A2-level performance is required to achieve score bands 3–4. A score of 5 is awarded for performances beyond A2 level.				

Speaking Task 2

Test	Aptis General	Component	Speaking	Task	Task 2

Features of the Task						
Skill focus	**Describing, expressing opinions, providing reasons and explanations in response to spoken questions**					
Task level (CEFR)	A1	A2	B1	B2	C1	C2
Task description	The candidate responds to three questions related to one picture prompt. The first question asks the candidate to describe a photograph. The candidate then responds to two questions related to a concrete and familiar topic represented in the photo. The candidate will be asked to give opinions and elaborate on the topic.					
Instructions to candidates	*Part two. In this part, I'm going to ask you to describe a picture. Then I will ask you two questions about it. You will have 45 seconds for each response. Begin speaking when you hear this sound (beep).*					
Presentation of rubric	Aural		Written		Visual non-verbal (e.g. photo)	
Response format	Q&A		Short turn		Long turn	
Planning time	None					
Delivery	Face-to-face	Telephone	Computer	Other		
Nature of input	Real time (face-to-face)	Real time (remote)	Pre-recorded input	No aural input		
	Unscripted	Guided	Semi-scripted	Scripted	N/A	
Nature of interaction	Interlocutor–Candidate (I–C)		Candidate–Candidate (C–C)			
	Candidate only (C)		Interlocutor–Candidate–Candidate			
Functions targeted	**Informational functions**		**Interactional functions**		**Managing interaction**	
	Providing personal information		Agreeing			
	Explaining opinions/preferences		Disagreeing		Initiating	
	Elaborating		Modifying/commenting		Changing topics	
	Justifying opinions		Asking for opinions		Reciprocating	
	Comparing		Persuading		Deciding	
	Speculating		Asking for information			
	Staging		Conversational repair			
	Describing		Negotiation of meaning			
	Summarising					
	Suggesting					
	Expressing preferences					

Features of the Input / Prompt										
Description	A single photograph of people engaged in a concrete, everyday activity. The recorded prompt asks 3 short questions related to the photograph: 1) describe the picture; 2) talk about an aspect of the photo relevant to the candidate's own context and experience; 3) elaborate by talking about the same topic in more general terms and providing an opinion with reasons and justification.									
Length of questions	Maximum of 15 words per questions									
Lexical level	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
Content knowledge	General									Specific
Cultural specificity	Neutral									Specific
Nature of information	Only concrete		Mostly concrete		Fairly abstract			Mainly abstract		
Relevant domain	Public		Occupational		Educational			Personal		
Topic	From topic list for A2/B1. The photograph will show several people engaged in an everyday, familiar activity. Appropriate questions will be about the activity and expand from asking the candidate to talk about similar activities in their own context to giving their opinions on the topic from a more general level.									

Features of the Expected Response	
Description	Short spoken responses to 3 questions. Candidate must provide sufficient content in response to at least 2 questions to achieve a rating of 3 (out of 5) for the task.
Length of response	Up to 45 seconds per question. Adequate responses will be beyond the single clause/sentence level.
Lexis/grammar	Demonstration of grammatical control at the B1 level necessary for a rating of 3 (out of 5) for the task. B1 lexis sufficient to respond adequately to all questions.
Rating scale for task	A task-specific holistic rating scale is used for the task. The rating scale is a 6-point scale from 0–5. A B1-level performance is required to achieve score bands 3–4. A score of 5 is awarded for performances beyond B1 level.

Speaking Task 3

Test	Aptis General	Component	Speaking	Task	Task 3

Features of the Task						
Skill focus	**Describing, comparing and contrasting, providing reasons and explanations to spoken questions**					
Task level (CEFR)	A1	A2	B1	B2	C1	C2
Task description	The candidate responds to 3 spoken questions about two photographs. The candidate is asked to describe, contrast and compare aspects of the photographs familiar to typical B1 Aptis General candidates. The candidate will be asked to compare aspects of the photos, give opinions, and provide reasons and explanations.					
Instructions to candidates	*Part three. In this part, I'm going to ask you to compare two pictures and I will ask you two questions about them. You will have 45 seconds for each response. Begin speaking when you hear this sound (beep).*					
Presentation of rubric	Aural		Written		Visual non-verbal (e.g. photo)	
Response format	Q&A		Short turn		Long turn	
Planning time	None					
Delivery	Face-to-face	Telephone	Computer	Other		
Nature of input	Real time (face-to-face)	Real time (remote)	Pre-recorded input	No aural input		
	Unscripted	Guided	Semi-scripted	Scripted	N/A	
Nature of interaction	Interlocutor–Candidate (I–C)		Candidate–Candidate (C–C)			
	Candidate only (C)		Interlocutor–Candidate–Candidate			
Functions targeted	**Informational functions**		**Interactional functions**		**Managing interaction**	
	Providing personal information		Agreeing			
	Explaining opinions/preferences		Disagreeing		Initiating	
	Elaborating		Modifying/commenting		Changing topics	
	Justifying opinions		Asking for opinions		Reciprocating	
	Comparing		Persuading		Deciding	
	Speculating		Asking for information			
	Staging		Conversational repair			
	Describing		Negotiation of meaning			
	Summarising					
	Suggesting					
	Expressing preferences					

Features of the Input / Prompt										
Description	Two photographs of scenes and/or activities which provide the basis for contrast and comparison on a topic/aspect familiar to B1-level candidates. The recorded prompt asks 3 short questions related to the photographs: 1) a description of both pictures; 2) to contrast and compare some aspect of the pictures; 3) to provide an opinion and/or express a preference in relation to the aspects already elaborated.									
Length of questions	Maximum of 15 words per questions									
Lexical level	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
Content knowledge	General						Specific			
Cultural specificity	Neutral						Specific			
Nature of information	Only concrete	Mostly concrete	Fairly abstract	Mainly abstract						
Relevant domain	Public	Occupational	Educational	Personal						
Topic	From topic list for B1. The photographs will show activities/and or scenes which can be compared and contrasted and will be familiar to a typical B1-level Aptis general candidate. The second question will focus on some aspect of the activities/scenes open to contrast and comparison, and the third question will extend the task by asking the candidate to express an opinion and/or preference in relation to some aspect of the photos.									

Features of the Expected Response	
Description	Short responses to 3 questions. Candidate must provide sufficient content in response to at least 2 questions to achieve a rating of 3 (out of 5) for the task.
Length of response	Up to 45 seconds per question. Adequate responses will be beyond the single clause/sentence level.
Lexis/grammar	Demonstration of grammatical control at the B1 level necessary for a rating of 3 (out of 5) for the task. B1 lexis sufficient to respond adequately to all questions.
Rating scale for task	A task-specific holistic rating scale is used for the task. The rating scale is a 6-point scale from 0–5. A B1-level performance is required to achieve score bands 3–4. A score of 5 is awarded for performances beyond B1 level.

Speaking Task 4

Test	Aptis General	Component	Speaking	Task	Task 4
Features of the Task					
Skill focus	**Integrating ideas regarding an abstract topic into a long turn. Giving opinions, justifying opinions, giving advantages and disadvantages.**				
Task level (CEFR)	A1	A2	B1	B2	C1	C2
Task description	The candidate plans a long turn integrating responses to a set of 3 questions related to a more abstract topic. The candidate speaks for two minutes to present his/her long-turn. The 3 questions expand in focus and cognitive demand (see features of the input/prompts below).				
Instructions to candidates	*Part four. In this part, I'm going to show you a picture and ask you three questions. You will have one minute to think about your answers before you start speaking. You will have two minutes to answer all three questions. Begin speaking when you hear this sound (beep). Look at the photograph.*				
Presentation of rubric	Aural	Written	Visual non-verbal (e.g. photo)		
Response format	Q&A	Short turn	Long turn		
Planning time	1 minute				
Delivery	Face-to-face	Telephone	Computer	Other	
Nature of input	Real time (face-to-face)	Real time (remote)	Pre-recorded input	No aural input	
	Unscripted	Guided	Semi-scripted	Scripted	N/A
Nature of interaction	Interlocutor–Candidate (I–C)	Candidate–Candidate (C–C)			
	Candidate only (C)	Interlocutor–Candidate–Candidate			
Functions targeted	**Informational functions**	**Interactional functions**	**Managing interaction**		
	Providing personal information	Agreeing			
	Explaining opinions/preferences	Disagreeing	Initiating		
	Elaborating	Modifying/commenting	Changing topics		
	Justifying opinions	Asking for opinions	Reciprocating		
	Comparing	Persuading	Deciding		
	Speculating	Asking for information			
	Staging	Conversational repair			
	Describing	Negotiation of meaning			
	Summarising				
	Suggesting				
	Expressing preferences				
Features of the Input / Prompt					
Description	Three questions. 1) Asks for a description of personal experience in relation to an abstract topic. 2) Asks for elaboration on the candidate's impression/opinion in relation to the topic. 3) Asks for a more objective discussion of the topic from the perspective of wider relevance to society/people in general. A photograph is provided for extra contextualisation but is not referred to in the questions.				
Length of questions	Maximum of 20 words per question				
Lexical level	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
Content knowledge	General					Specific				
Cultural specificity	Neutral					Specific				
Nature of information	Only concrete	Mostly concrete	Fairly abstract	Mainly abstract						
Relevant domain	Public	Occupational	Educational	Personal						
Topic	From topic list for B2.									
Features of the Expected Response										
Description	A long turn of 2 minutes. Candidate must provide a coherent and cohesive long turn which deals with at least 2 questions to achieve a rating of 3 (out of 5) for the task.									
Length of response	Up to 2 minutes for the entire long turn. Adequate length for B2-level performance will generally require the candidate to speak for the full two minutes or most of the full two minutes.									
Lexis/grammar	Demonstration of grammatical control at the B2 level necessary for a rating of 3 (out of 5) for the task. B2 lexis sufficient to respond adequately to all questions.									
Rating scale for task	A task-specific holistic rating scale is used for the task. The rating scale is a 7-point scale from 0–6. A B2-level performance is required to achieve score bands 3–4. A score of 5 or 6 is awarded for performances beyond B2 level, with a 5 describing performance equivalent to a C1 level, and 6 for performances at a C2 level.									

Appendix F: Task specifications for Aptis General Writing component

Writing Task 1

Test	Aptis General	Component	Writing	Task	Task 1
Features of the Task					
Skill focus	**Writing at the word level. Simple personal information on a form.**				
Task level (CEFR)	A1	A2		B2	C1	C2
Task description	The candidate completes a form by filling in some basic personal information. All responses are at the word-level, inputting information such as name, birthdate, etc. in a form. Each form will consist of five categories of information with a total of 9 gaps in a consistent format (see features of the response below).				
Instructions to candidates	The instructions will clearly identify the purpose of the form to be completed. The following is an example only, and other purposeful activities within the relevant domains which could support the kinds of writing required in all 4 tasks should also be developed: *You want to join a travel club. Fill in the form.*				
Presentation of rubric	Aural	Written	Other non-verbal (e.g. photo)		
Time for task	50 minutes for entire Writing test. No time limit is set for individual tasks. (3 minutes recommended for Task 1).				
Delivery	Pen and paper	Computer			
Response format	Word completion	Gap-filling	Form filling	Short answer	Continuous writing
Intended genre	Simple form for providing personal details				
Writer / intended reader relationship	The reader will not be known to the writer. The writing is transactional in nature and the reader is understood to be anyone associated with processing the form for the intended function of the activity in the task setting.				
Discourse mode	Descriptive	Narrative	Expository	Argumentative	Instructive
Domain	Public	Occupational	Educational	Personal	
Nature of task	Knowledge telling	Knowledge transformation			
Functions targeted	Providing personal information (Based on British Council EQUALS Core Inventory)				
Features of the Input / Prompt					
Description	Short form. Categories to be filled are clearly labelled on the left hand side of the form followed by space for inputting necessary information by the candidate.				
Number of categories	There will be five categories: (a) *full name*, (b) *country (where you live),* (c) *date of birth*, (d) *first language* or *job*, (e) final category asks for list of 3 things relevant to the overall activity of the task setting (e.g. interests, favourite subjects, etc.).				
Number of gaps	(a) 1, (b) 1, (c) 3 (day, month, year), (d) 1, (e) 3 (the candidate will be asked to list 3 different pieces of information for this category, e.g. 3 interests, or 3 modes of travel)				
Lexical level	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
Content knowledge	General				Specific
Cultural specificity	Neutral				Specific
Nature of information	Only concrete	Mostly concrete	Fairly abstract	Mainly abstract	
Relevant domain	Public	Occupational	Educational	Personal	
Information targeted	Personal information which is easily recoverable from memory and which an A1-level candidate is expected to be able to communicate. At least one category should target numbers and/or dates.				
Features of the Expected Response					
Description	9 short gaps which can be filled by 1–2 word responses.				
Length of response	Each gap can be filled by 1–2 word responses.				
Lexis/grammar	K1 level lexis sufficient to complete task. Some personal information may not be on the K1 list, such as first language or proper nouns for home town, etc., but is still appropriate if it is the kind of very familiar, personal information which is required in everyday situations targeted by the task.				
Rating scale for task	A task-specific rating scale is used for the task. The rating scale is a 6-point scale from 0–5. Marks are awarded for correctly supplied information as specified in the rating scheme. Spelling, capitalisation, punctuation, and formatting of dates and numbers are specified in the marking scheme where appropriate.				

Writing Task 2

Test	Aptis General	Component	Writing	Task	Task 2
Features of the Task					
Skill focus	**Short written description of concrete, personal information at the sentence level.**				
Task level (CEFR)	A1	A2		B2	C1	C2
Task description	The candidate continues filling in information on a form. The task setting and topic are related to the same purpose as the form used in part 1. The candidate must write a short response using sentence-level writing to provide personal information in response to a single written question.					
Instructions to candidates	The instructions will clearly identify the purpose of the form to be completed. The following is an example only, and other kinds of follow-up questions appropriate to the setting and the A2-level targeted should be developed: *You are a new member of the travel club. Write in sentences. Use 20–30 words.*					
Presentation of rubric	Aural	Written	Other non-verbal (e.g. photo)			
Time for task	50 minutes for entire Writing test. No time limit is set for individual tasks. (7 minutes recommended for Task 2).					
Delivery	Pen and paper	Computer				
Response format	Word completion	Gap-filling	Form filling	Short answer	Continuous writing	
Intended genre	Section of a simple form for providing personal details					
Writer / intended reader relationship	The reader will not be known to the writer. The writing is transactional in nature and the reader is understood to be anyone associated with processing the form for the intended function of the activity in the task setting.					
Discourse mode	Descriptive	Narrative	Expository	Argumentative	Instructive	
Domain	Public	Occupational	Educational	Personal		
Nature of task	Knowledge telling	Knowledge transformation				
Functions targeted	Describing (people, places, job), describing likes/dislike/ interests, describing habits and routines, describing past experiences (Based on British Council EQUALS Core Inventory)					
Features of the Input / Prompt						
Description	Short sentence specifying what kind of information the candidate is expected to provide.					
Length	10–15 words					
Lexical level	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
Content knowledge	General			Specific		
Cultural specificity	Neutral			Specific		
Nature of information	Only concrete	Mostly concrete	Fairly abstract	Mainly abstract		
Relevant domain	Public	Occupational	Educational	Personal		
Information targeted	The information targeted would be concrete, everyday, and familiar information about the candidate, the candidate's personal experiences or surroundings, occupation, everyday activities etc.					
Features of the Expected Response						
Description	A short constructed response. Responses need to be structured as sentences to receive a rating of 3 or more (out of 5).					
Length of response	20–30 words					
Lexis/grammar	K1–K2 level lexis sufficient to complete task. Response needs to demonstrate control of A2-level grammar, writing at the sentence level.					
Rating scale for task	A task-specific holistic rating scale is used for the task. The rating scale is a 6-point scale from 0–5. An A2-level performance is required to achieve score bands 3–4. A score of 5 is awarded for performances beyond A2 level.					

Writing Task 3

Test	Aptis General	Component	Writing	Task	Task 3
Features of the Task					
Skill focus	**Interactive writing. Responding to a series of written questions with short paragraph-level responses.**				
Task level (CEFR)	A1	A2	B1	B2	C1	C2
Task description	The candidate responds interactively to three separate questions. Each response requires a short paragraph-level response. The questions are presented as if the candidate is writing on an internet forum or social network site. The task setting and topic are related to the same background activity used in parts 1 & 2.					
Instructions to candidates	The instructions identify the setting for the interaction and person or persons with whom the candidate is interacting. The following is an example only, and other kinds of follow-up questions appropriate to the setting and the B1-level targeted should be developed: *You are a member of a travel club. Talk to other members in the travel club chat room. Talk to them using sentences. Use 30–40 words per answer.*					
Presentation of rubric	Aural	Written	Other non-verbal (e.g. photo)			
Time for task	50 minutes for Writing test. No time limit is set for individual tasks. (10 minutes recommended for Task 1)					
Delivery	Pen and paper	Computer				
Response format	Word completion	Gap-filling	Form filling	Short answer	Continuous writing	
Intended genre	Interaction in a social-media context. The context for interaction may be within the public, occupational, or educational domains, reflecting real-life situations in which interactive, information-exchange forums might be used, but which do not require specialist knowledge or experience (e.g. students in an online course discussing course options, favourite subjects and educational features of the candidate's own educational context).					
Writer/intended reader relationship	The reader will be specified. The reader is not personally known to the candidate but is a participant in the same public/occupational/educational domain. Given the nature of the social media task, the message will be accessible to others.					
Discourse mode	Descriptive	Narrative	Expository	Argumentative	Instructive	
Domain	Public	Occupational	Educational	Personal		
Nature of task	Knowledge telling	Knowledge transformation				
Functions targeted	Describing (people, places, job), describing likes/dislike/ interests, describing habits and routines, describing past experiences, describing feelings, emotions, attitudes, describing hopes and plans, expressing opinions, expressing agreement/disagreement					
Features of the Input / Prompt						
Description	Series of 3 prompts phrased as posts requesting information from the candidate by a member of the interactive forum.					
Length of posts	Each post requesting information should be in the form of 1–3 short sentences. Maximum length of a post is 25–30 words, with no one sentence more than 13–15 words.					
Lexical level	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
Grammatical level	A2 Grammatical exponents (See Guidelines on Adhering to Grammatical Level)					
Content knowledge	General				Specific	
Cultural specificity	Neutral				Specific	
Nature of information	Only concrete	Mostly concrete	Fairly abstract	Mainly abstract		
Relevant domain	Public	Occupational	Educational	Personal		
Information targeted	The information targeted should be familiar to the candidate and may include talking about the candidate's personal experiences, plans, etc. One question should ask the candidate to describe some aspect of the candidate's own context from a wider a perspective than the candidate's personal experience (describing features of the educational or working context in the candidate's country, subjects typically studied, etc.).					
Features of the Expected Response						
Description	A series of 3 short constructed responses. Each response needs to be structured as sentences, and the candidate must respond adequately to at least 2 questions to receive a rating of 3 or more (out of 5).					
Length of response	30–40 words per response					
Lexis/grammar	K1–K3 level lexis sufficient to complete task. Response needs to demonstrate control of B1-level grammar, writing at the short paragraph level.					
Rating scale for task	A task-specific holistic rating scale is used for the task. The rating scale is a 6-point scale from 0–5. A B1-level performance is required to achieve score bands 3–4. A score of 5 is awarded for performances beyond B1 level.					

Writing Task 4

Test	Aptis General	Component	Writing	Task	Task 4
Features of the Task					
Skill focus	**Integrated writing task requiring longer paragraph level writing in response to two emails. Use of both formal/informal registers required.**				
Task level (CEFR)	A1 / A2 / B1 / **B2** / C1 / C2				
Task description	The candidate writes two emails in response to the task prompt which contains a short letter/notice. The first email response is an informal email to a friend regarding the information in the task prompt. The second is a more formal email to an unknown reader connected to the information (management, customer services, etc.)				
Instructions to candidates	The instructions will clearly identify the purpose by presenting a transactional email from the organisation which provides the background setting for all tasks (school offering online course, management of company, management of club/business etc.). The email will present a problem/issue/offer/opportunity which the candidate is expected to discuss in two different registers. The following is an example only: *You are a member of a travel club. You receive this email from the club: (text of short transactional email message). Write an email to your friend about your feelings and what you plan to do. Write about 50 words. Write an email to the secretary of the club. Write about your feelings and what you would like to do. Write 120–150 words.*				
Presentation of rubric	Aural	**Written**	Other non-verbal (e.g. photo)		
Time for task	50 minutes for Writing test. No time limit is set for individual tasks. (10 minutes recommended for first email, and 20 minutes for the second email).				
Delivery	Pen and paper	**Computer**			
Response format	Word completion / Gap-filling / Form filling / Short answer / **Continuous writing**				
Intended genre	Emails, one informal, the other formal				
Writer/intended reader relationship	The readers are specified. The first reader will be known to the candidate as a participant in the same background activity as Tasks 1, 2, 3 (colleague, student studying on same online course, member of same club, etc.). Although the reader of the first email is known and the register is informal, the reader/writer relationship is defined by their roles as participants in the same activity in the public/occupational/educational domain. The intended reader of the second email will be specified but may or may not be personally known to the writer.				
Discourse mode	Descriptive / Narrative / **Expository** / **Argumentative** / Instructive				
Domain	**Public** / **Occupational** / **Educational** / Personal				
Nature of task	Knowledge telling / **Knowledge transformation**				
Functions targeted	Expressing opinions, giving reasons and justifications, describing hopes and plans, giving precise information, expressing abstract ideas, expressing certainty/probability/doubt, generalising and qualifying, synthesising, evaluating, speculating and hypothesising, expressing opinions tentatively, expressing shades of opinion, expressing agreement/ disagreement, expressing reaction, e.g. indifference, developing an argument systematically, conceding a point, emphasising a point/feeling/issue, defending a point of view persuasively, complaining, suggesting (based on British Council Equals Core Inventory)				
Features of the Input / Prompt					
Description	A transactional email message is presented as the starting point for both email responses to be produced. A separate instruction of 1–2 sentences is given for each email response. The instructions will specify the intended reader and the purpose/function of the email (complaining, suggesting alternatives, giving advice, etc.).				
Length of input email	50–80 words				
Lexical level	K1 / K2 / K3 / K4 / K5 / K6 / K7 / K8 / K9 / K10				
Content knowledge	**General** / Specific				
Cultural specificity	**Neutral** / Specific				
Nature of information	Only concrete / **Mostly concrete** / **Fairly abstract** / Mainly abstract				
Relevant domain	**Public** / **Occupational** / **Educational** / Personal				
Information targeted	The information will be relevant to eliciting more complex and abstract functions described above.				
Features of the Expected Response					
Description	Two separate emails, one in an informal register, one in a formal register.				
Length of response	Approximately 50 words for the first email, 120–150 words for the second email.				
Lexis/grammar	K4–K5 lexis will be sufficient to complete both emails adequately. Responses must show control of B2-level grammar and cohesion and coherence across longer continuous writing texts.				
Rating scale for task	A task-specific holistic rating scale is used for the task. The rating scale is a 7-point scale from 0–6. A B2-level performance is required to achieve score bands 3–4. A score of 5 or 6 is awarded for performances beyond B2 level, with a 5 describing performance equivalent to a C1 level, and 6 for performances at a C2 level.				

Appendix G: List of topics (offered as general guidelines only)

This is a generic list of possible topics covering a range of proficiency levels. The topics have been developed considering a broad range of potential Target Language Use domains for general English use situations in both EFL and ESL contexts. At A1, appropriate topics focus on everyday, familiar activities and aspects of daily life. A wider range of activities and more abstract topics become relevant as the levels increase.

Topic	A1	A2	B1	B2
Architecture			■	■
Arts (art, dance, film, literature, music)			■	■
Biographies		■	■	■
Business, finance, industry			■	■
Culture and customs		■	■	■
Daily life	■	■	■	■
Descriptions of buildings	■	■	■	
Descriptions of places (towns, cities, locations)	■	■	■	
Descriptions of people (appearance, personality)	■	■	■	
Dreams and future plans		■	■	■
Education — college life		■	■	■
Education — school life		■	■	■
Education — social topic			■	■
Education — training and learning		■	■	■
Environmental issues			■	■
Food and drink	■	■	■	
Health and medicine — social topic			■	■
Health and injuries — personal health	■	■	■	■
History and archaeology			■	■
Humanitarian and volunteer activities			■	■
Leisure and entertainment	■	■	■	■
Media			■	■
Personal finances		■	■	■
Pets	■	■	■	
Plants, animals, nature	■	■	■	■
Politics and government			■	■
Public safety — accidents and natural disasters			■	■
Public safety — crime			■	■
Relationships and family	■	■	■	■
Science and technology		■	■	■
Shopping and obtaining services	■	■	■	■
Social trends			■	■
Sports		■	■	■
Transportation and asking for directions	■	■	■	■
Travel and tourism	■	■	■	■
Weather	■	■	■	
Work and job related		■	■	■

Appendix H: Rating scales for Speaking and Writing

The following examples provide descriptions of the performance expected at each score point band in the task-specific rating scales used for rating the Speaking and Writing components. The rating scales are described further in Section 3.3.3.3 of the manual. Each scale is task-specific. The 3- and 4-point score bands for each scale describe the target-level performance at the proficiency level targeted by that task.

Speaking Task 1

Areas assessed: task fulfilment / topic relevance, grammatical range & accuracy, vocabulary range & accuracy, pronunciation, fluency.

5 **B1 (or above)**	Likely to be above A2 level.
4 **A2.2**	Responses to all **three** questions are on topic and show the following features • Some simple grammatical structures used correctly but basic mistakes systematically occur. • Vocabulary is sufficient to respond to the questions, although inappropriate lexical choices are noticeable. • Mispronunciations are noticeable and frequently place a strain on the listener. • Frequent pausing, false starts and reformulations but meaning is still clear.
3 **A2.1**	Responses to **two** questions are on topic and show the following features • Some simple grammatical structures used correctly but basic mistakes systematically occur. • Vocabulary is sufficient to respond to the questions, although inappropriate lexical choices are noticeable. • Mispronunciations are noticeable and frequently place a strain on the listener. • Frequent pausing, false starts and reformulations but meaning is still clear.
2 **A1.2**	Responses to at least **two** questions are on topic and show the following features • Grammatical structure is limited to words and phrases. Errors in basic patterns and simple grammar structures impede understanding. • Vocabulary is limited to very basic words related to personal information. • Pronunciation is mostly unintelligible except for isolated words. • Frequent pausing, false starts and reformulations impede understanding.
1 **A1.1**	Response to **one** question is on topic and shows the following features • Grammatical structure is limited to words and phrases. Errors in basic patterns and simple grammar structures impede understanding. • Vocabulary is limited to very basic words related to personal information. • Pronunciation is mostly unintelligible except for isolated words. • Frequent pausing, false starts and reformulations impede understanding.
0 **A0**	• No meaningful language or all responses are completely off-topic (e.g. memorised script, guessing).

Speaking Tasks 2 and 3

Areas assessed: task fulfilment / topic relevance, grammatical range & accuracy, vocabulary range & accuracy, pronunciation, fluency and cohesion.

5 **B2 (or above)**	Likely to be above B1 level.
4 **B1.2**	Responses to all **three** questions are on topic and show the following features • Control of simple grammatical structures. Errors occur when attempting complex structures. • Sufficient range and control of vocabulary for the task. Errors occur when expressing complex thoughts. • Pronunciation is intelligible but inappropriate mispronunciations put an occasional strain on the listener. • Some pausing, false starts and reformulations. • Uses only simple cohesive devices. Links between ideas are not always clearly indicated.
3 **B1.1**	Responses to **two** questions are on topic and show the following features • Control of simple grammatical structures. Errors occur when attempting complex structures. • Sufficient range and control of vocabulary for the task. Errors occur when expressing complex thoughts. • Pronunciation is intelligible but inappropriate mispronunciations put an occasional strain on the listener. • Some pausing, false starts and reformulations. • Uses only simple cohesive devices. Links between ideas are not always clearly indicated.
2 **A2.2**	Responses to at least **two** questions are on topic and show the following features • Uses some simple grammatical structures correctly but systematically makes basic mistakes. • Vocabulary will be limited to concrete topics and descriptions. Inappropriate lexical choices for the task are noticeable. • Mispronunciations are noticeable and put a strain on the listener. • Noticeable pausing, false starts and reformulations. • Cohesion between ideas is limited. Responses tend to be a list of points.
1 **A2.1**	Response to **one** question is on topic and shows the following features • Uses some simple grammatical structures correctly but systematically makes basic mistakes. • Vocabulary will be limited to concrete topics and descriptions. Inappropriate lexical choices for the task are noticeable. • Mispronunciations are noticeable and put a strain on the listener. • Noticeable pausing, false starts and reformulations. • Cohesion between ideas is limited. Responses tend to be a list of points.
0	• Performance below A2, or no meaningful language or the responses are completely off-topic (e.g. memorised script, guessing).

Speaking Task 4

Areas assessed: task fulfilment / topic relevance, grammatical range & accuracy, vocabulary range & accuracy, pronunciation, fluency and cohesion.

6 **C2**	Likely to be above C1 level.
5 **C1**	Response addresses all **three** questions and is well-structured. • Uses a range of complex grammar constructions accurately. Some minor errors occur but do not impede understanding. • Uses a range of vocabulary to discuss the topics required by the task. Some awkward usage or slightly inappropriate lexical choices. • Pronunciation is clearly intelligible. • Backtracking and reformulations do not fully interrupt the flow of speech. • A range of cohesive devices are used to clearly indicate the links between ideas.
4 **B2.2**	Responses to all **three** questions are on topic and show the following features • Some complex grammar constructions used accurately. Errors do not lead to misunderstanding. • Sufficient range of vocabulary to discuss the topics required by the task. Inappropriate lexical choices do not lead to misunderstanding. • Pronunciation is intelligible. Mispronunciations do not put a strain on the listener or lead to misunderstanding. • Some pausing while searching for vocabulary but this does not put a strain on the listener. • A limited number of cohesive devices are used to indicate the links between ideas.
3 **B2.1**	Responses to **two** questions are on topic and show the following features • Some complex grammar constructions used accurately. Errors do not lead to misunderstanding. • Sufficient range of vocabulary to discuss the topics required by the task. Inappropriate lexical choices do not lead to misunderstanding. • Pronunciation is intelligible. Mispronunciations do not put a strain on the listener or lead to misunderstanding. • Some pausing while searching for vocabulary but this does not put a strain on the listener. • A limited number of cohesive devices are used to indicate the links between ideas.
2 **B1.2**	Responses to at least **two** questions are on topic and show the following features • Control of simple grammatical structures. Errors occur when attempting complex structures. • Limitations in vocabulary make it difficult to deal fully with the task. • Pronunciation is intelligible but occasional mispronunciations put an occasional strain on the listener. • Noticeable pausing, false starts, reformulations and repetition. • Uses only simple cohesive devices. Links between ideas are not always clearly indicated.
1 **B1.1**	Response to **one** question is on topic and shows the following features • Control of simple grammatical structures. Errors occur when attempting complex structures. • Limitations in vocabulary make it difficult to deal fully with the task. • Pronunciation is intelligible but occasional mispronunciations put an occasional strain on the listener. • Noticeable pausing, false starts, reformulations and repetition. • Uses only simple cohesive devices. Links between ideas are not always clearly indicated.
0 **A1/A2**	Performance not sufficient for B1, or no meaningful language, or the responses are completely off-topic (memorised or guessing).

Writing Task 2

Areas assessed: task fulfilment / topic relevance, grammatical range & accuracy, punctuation, vocabulary range & accuracy, cohesion.

5 **B1 (or above)**	Likely to be above A2 level.
4 **A2.2**	• On topic. • Uses simple grammatical structures to produce writing at the sentence level. Errors with basic structures common. Errors do not impede understanding of the response. • Mostly accurate punctuation and spelling. • Vocabulary is sufficient to respond to the question(s). • Some attempts at using simple connectors and cohesive devices to link sentences.
3 **A2.1**	• On topic • Uses simple grammatical structures to produce writing at the sentence level. Errors with basic structures common. Errors impede understanding in parts of the response. • Punctuation and spelling mistakes are noticeable. • Vocabulary is mostly sufficient to respond to the question(s) but inappropriate lexical choices are noticeable. • Response is a list of sentences with no use of connectors or cohesive devices to link sentences.
2 **A1.2**	• Not fully on topic • Grammatical structure is limited to words and phrases. Errors in basic patterns and simple grammar structures impede understanding. • Little or no use of accurate punctuation. Spelling mistakes common. • Vocabulary is limited to very basic words related to personal information and is not sufficient to respond to the question(s). • No use of cohesion.
1 **A1.1**	• Response limited to a few words or phrases. • Grammar and vocabulary errors so serious and frequent that meaning is unintelligible.
0 **A0**	No meaningful language or all responses are completely off-topic (e.g. memorised script, guessing).

Writing Task 3

Areas assessed: task fulfilment / topic relevance, punctuation, grammatical range & accuracy, vocabulary range & accuracy, cohesion.

5 **B2 (or above)**	Likely to be above the B1 level.
4 **B1.2**	Responses to all **three** questions are on topic and show the following features • Control of simple grammatical structures. Errors occur when attempting complex structures. • Punctuation and spelling mostly accurate. Errors do not impede understanding. • Vocabulary is sufficient to respond to the questions. • Uses simple cohesive devices to organise responses as a linear sequence of sentences.
3 **B1.1**	Responses to **two** questions are on topic and show the following features • Control of simple grammatical structures. Errors occur when attempting complex structures. • Punctuation and spelling mostly accurate. Errors do not impede understanding. • Vocabulary is sufficient to respond to the questions. • Uses simple cohesive devices to organise responses as a linear sequence of sentences.
2 **A2.2**	Responses to at least **two** questions are on topic and show the following features • Uses simple grammatical structures to produce writing at the sentence level. Errors with simple structures common and sometimes impede understanding. • Punctuation and spelling mistakes are noticeable. • Vocabulary is not sufficient to respond to the question(s). Inappropriate lexical choices are noticeable and sometimes impede understanding. • Responses are lists of sentences and not organised as cohesive texts.
1 **A2.1**	Response to **one** question is on topic and shows the following features • Uses simple grammatical structures to produce writing at the sentence level. Errors with simple structures common and sometimes impede understanding. • Punctuation and spelling mistakes are noticeable. • Vocabulary is not sufficient to respond to the question(s). Inappropriate lexical choices are noticeable and sometimes impede understanding. • Responses are lists of sentences and not organised as cohesive texts.
0	Performance below A2, or no meaningful language or the responses are completely off-topic (e.g. memorised script, guessing).

Writing Task 4

Areas assessed: task fulfilment & register, grammatical range & accuracy, vocabulary range & accuracy, cohesion.

6 **C2**	Likely to be above C1 level.
5 **C1**	Response shows the following features • Response on topic and task fulfilled in terms of appropriateness of register. Two clearly different registers. • Range of complex grammar constructions used accurately. Some minor errors occur but do not impede understanding. • Range of vocabulary used to discuss the topics required by the task. Some awkward usage or slightly inappropriate lexical choices. • A range of cohesive devices is used to clearly indicate the links between ideas.
4 **B2.2**	Response on topic and task fulfilled in terms of appropriateness of register: appropriate register used consistently in both responses. Response shows the following features • Some complex grammar constructions used accurately. Errors do not lead to misunderstanding. • Minor errors in punctuation and spelling occur but do not impede understanding. • Sufficient range of vocabulary to discuss the topics required by the task. Inappropriate lexical choices do not lead to misunderstanding. • A limited number of cohesive devices are used to indicate the links between ideas.
3 **B2.1**	Response partially on topic and task partially fulfilled in terms of appropriateness of register: appropriate register used consistently in one response. Response shows the following features • Some complex grammar constructions used accurately. Errors do not lead to misunderstanding. • Minor errors in punctuation and spelling occur but do not impede understanding. • Sufficient range of vocabulary to discuss the topics required by the task. Inappropriate lexical choices do not lead to misunderstanding. • A limited number of cohesive devices are used to indicate the links between ideas.
2 **B1.2**	Response partially on topic and task not fulfilled in terms of appropriateness of register: appropriate register not used consistently in either response. Response shows the following features • Control of simple grammatical structures. Errors occur when attempting complex structures. • Punctuation and spelling is mostly accurate. Errors do not impede understanding. • Limitations in vocabulary make it difficult to deal fully with the task. Errors impede understanding in parts of the text. • Uses only simple cohesive devices. Links between ideas are not always clearly indicated.
1 **B1.1**	Response not on topic and task not fulfilled in terms of appropriateness of register. No evidence of awareness of register. Response shows the following features • Control of simple grammatical structures. Errors occur when attempting complex structures. • Punctuation and spelling is mostly accurate. Errors do not impede understanding. • Limitations in vocabulary make it difficult to deal fully with the task. Errors impede understanding in most of the text. • Uses only simple cohesive devices. Links between ideas are not always clearly indicated.
0 **A1/A2**	Performance below B1, or no meaningful language or the responses are completely off-topic (e.g. memorised script, guessing).

Appendix I: Sample score reports

BRITISH COUNCIL

Aptis
Forward thinking
English testing

Candidate Report

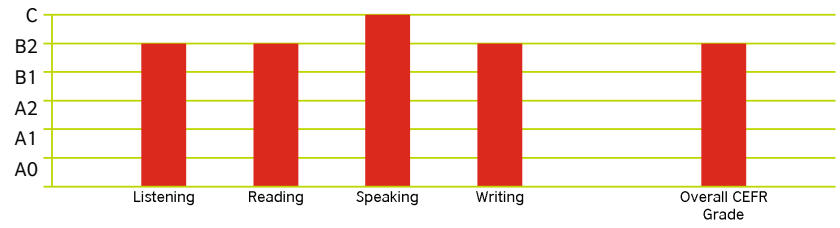
Candidate Name: **M Mike** Test Date: **01/07/2014**

Organization: **Aptis Control** Test Package: **4 Skills Package**

Scale Score

Skill Name	Skill Score
Listening	32/50
Reading	38/50
Speaking	50/50
Writing	42/50
Final Scale Score	**162/200**
Grammar & Vocab	**50/50**

CEFR Skill Profile



Please turn over for CEFR Skill Descriptors

www.britishcouncil.org

BRITISH COUNCIL Aptis Forward thinking English testing

CEFR Skill Descriptors

Listening

A0	Not enough to allow for any meaningful inferences about the candidate's ability.
A1	Can follow speech which is very slow and carefully articulated, with long pauses for him/her to assimilate meaning.
A2	Can understand enough to be able to meet needs of a concrete type provided speech is clearly and slowly articulated.
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.
B2	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.
C	Has no difficulty in understanding any kind of spoken language, whether live or broadcast, delivered at fast native speed.

Reading

A0	Not enough to allow for any meaningful inferences about the candidate's ability.
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.
A2	Can understand short, simple texts on familiar matters of a concrete type which consist of high frequency everyday or job-related language.
B1	Can read straightforward factual texts on subjects related to his/her field and interest with a satisfactory level of comprehension.
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.
C	Can understand and interpret critically virtually all forms of the written language.

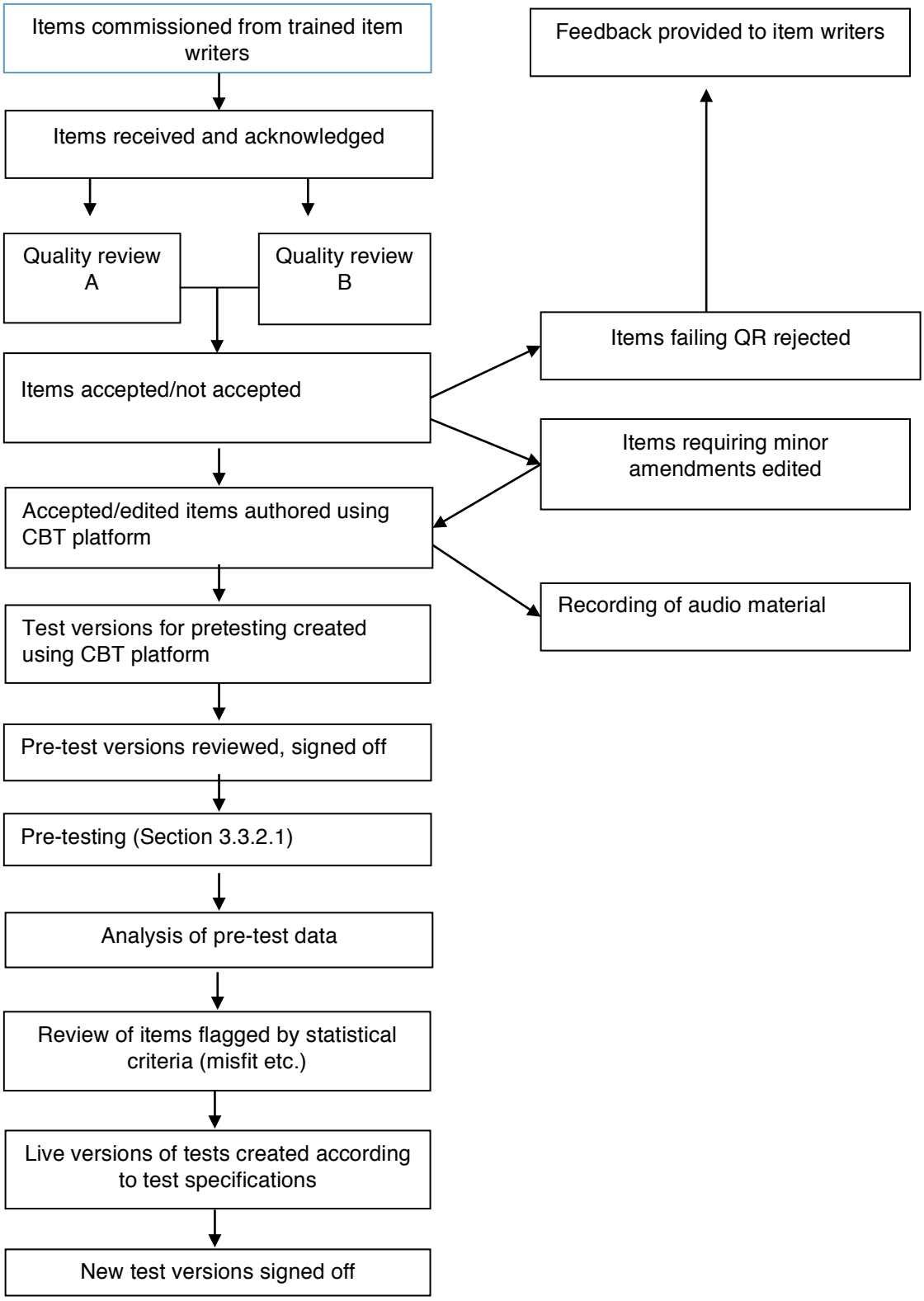
Speaking

A0	Not enough to allow for any meaningful inferences about the candidate's ability.
A1	Can produce simple descriptions on mainly personal topics.
A2	Can give a simple description or presentation of people, living or working conditions, daily routines likes/dislikes, etc. as a short series of simple phrases and sentences linked into a list
B1	Can reasonably fluently sustain a straightforward description of one of a variety of subjects within his/her field of interest, presenting it as a linear sequence of points.
B2	Can give clear, systematically developed descriptions and presentations on a wide range of subjects related to his/her field of interest, with appropriate highlighting of significant points, and relevant supporting detail.
C	Can produce clear, smoothly flowing well-structured speech with an effective logical structure which helps the recipient to notice and remember significant points.

Writing

A0	Not enough to allow for any meaningful inferences about the candidate's ability.
A1	Can write simple isolated phrases and sentences.
A2	Can write a series of simple phrases and sentences linked with simple connectors like 'and', 'but' and 'because'.
B1	Can write straightforward connected texts on a range of familiar subjects within his field of interest, by linking a series of shorter discrete elements into a linear sequence.
B2	Can write clear, detailed texts on a variety of subjects related to his/her field of interest and shows an ability to use different registers within written texts.
C	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.

Appendix J: Flow chart of the item and test production cycle



Glossary

Analytic scale	Analytic score scales are a set of separate rating scales used to rate a constructed response task / item, with each scale focusing on one specific aspect of performance. Analytic scales are often contrasted with holistic scales (see holistic scale).
Candidate	An individual test-taker.
CEFR	The Common European Framework of Reference for Languages: Learning, Teaching, Assessment (Council of Europe, 2001).
Certificated test	A test that has an official certification process. The certificate issued to test-takers can be used as official proof of the proficiency level demonstrated by the test-taker for the skill or ability which the examination tests. Test results are thus recognised for use beyond one specific organisation or context.
Component	Component is used here to refer to a distinctly separate *part* of an overall assessment product, which has its own scoring, time limits, etc., and for which a score and/or CEFR level is reported. There are 5 components in Aptis General (the Core, Reading, Listening, Speaking and Writing). In general usage, components are also referred to as different papers or tests (e.g. the listening paper, or the listening test).
Constructed response	The candidate must produce the response from their own linguistic resources, for example, write one or more words to respond to a writing task, or create an oral response to respond to a speaking task. (For language proficiency tests, these are mostly associated with productive skills, speaking and writing.)
Distractor	Incorrect option for selected response (multiple choice response type items).
Holistic scale	A single score scale used to rate a constructed response task / item. For example, a speaking task may be rated using a holistic rating scale of 0–5, with each score band containing a description of the performance necessary to achieve that score. The performance at each band may contain a number of dimensions (for example, in order to achieve a score of 5, a candidate may need to use certain vocabulary, have a certain level of grammar, and certain level of pronunciation). Holistic rating scales are often contrasted with analytic rating scales, in which each of those dimensions (vocabulary, etc.) is scored separately on its own scale.
Item	Each stand-alone, single response by the test-taker which can be marked correct/incorrect or given a single rating. An item is the minimum level of quantitative response data scored. An item can be a discrete selected response item (e.g., a single question followed by four response alternatives for which the candidate selects only one response which is scored correct or incorrect, a single gap in a gap fill task, a label that has to be matched to the right paragraph or correct illustration, etc.). An item may also be a constructed response item, for example, an answer to a question in a speaking test that is scored using a rating scale, or a single long response, for example an essay response to a single essay prompt. A group of items may be grouped together into a task, but each item will still be scored separately. All test analysis for score reporting and test validation requires quantitative response data to be captured at the item level.
Key	The intended correct answer for scoring.
Option	One of a set of options provided to candidates for selected-response items in which a test-taker selects the correct option (or options) from a list of choices.
Package	A test package refers to the particular combination of components to be used in a particular administration by a particular group of test-takers. Aptis General has 5 separate components: Core (Grammar and Vocabulary); Reading; Listening; Speaking; and Writing. The components can be combined in different ways to form specified *test packages*: for example, a *speaking package* contains the Core component + the Speaking component, while a Reading and Listening package contains the Core component + Reading + Listening, etc. A full package is also referred to as a four-skills package, as it contains components focusing each of the four main skills, listening, reading, speaking, and writing, in addition to the Core component which focuses on language knowledge.

Rasch	A form of statistical analysis within the family of item response theory (IRT) measurement models. Rasch analysis is mathematically equivalent to the one-parameter model in IRT. Rasch uses what is called the simple logistic model to estimate the ability of a test-taker and the difficulty of a test item on a common scale of measurement which uses units referred to as logits.
Rater	The person who scores a test-taker's response to a test task or item using a specified scoring procedure. Raters in the Aptis test system are also referred to as examiners. All raters are trained and they use an explicit rating scale.
Rating scale	A scoring scale for constructed response items that are scored according to a defined set of criteria. Rating scales can have different numbers of categories. For example, a speaking task might be scored on a rating scale of 0–3 points, or on a scale of 0–5 points. Each score point (or score band) will usually be defined by descriptors which define the type of performance appropriate for each score. Two types of rating scale are commonly used: analytic scales and holistic scales (see entries under *analytic scale, holistic scale* for definitions).
Response format	The method used by a test-taker to respond to a test task or item. Two broad distinctions are commonly made, referred to as selected-response formats and constructed-response formats.
Rubric	The set of instructions given to a test-taker for a specific test task or item.
Selected response	The options are provided and the candidate must select the right option, or manipulate the option provided in a particular way. For language proficiency tests, these are mostly associated with receptive skills (e.g. language knowledge, reading, listening, etc.). Selected response formats are not limited to multiple-choice question formats, and include (but are not limited to), multiple choice gap-fill or sentence completion, matching, multiple matching, and re-ordering formats.
Specifications	A set of detailed documents that clearly describe the design and structure of test tasks and tests. Specifications for Aptis General have been derived using the socio-cognitive model of language test development and validation. Two types of specifications are referred to in this manual: *task* specifications and *test* specifications. *Task specifications* describe all elements of a test task necessary to create different forms of the same task which are comparable in terms of key features. *Test specifications* refer to the overall design template for a full test, specifying the number of tasks and items to be included, the scoring system, the time constraints, etc. Both types of specifications are used by the production team to ensure the comparability of tasks and versions of the same component.
Target	The intended correct answer for scoring.
Task	A task combines one set of instructions with the input to be processed and the activity or activities to be carried out by the candidate. A task has one or more items based on the same input text or texts. Examples include: a reading text, graph or illustration which comes with a set of related reading comprehension questions; a listening input text followed by an activity in which candidates match participants in the input text with the opinions expressed by each participant; an activity designed to elicit a constructed response performance, e.g. responding to one or more spoken questions about an illustration in a speaking task, writing a constructed response on a given topic for a writing task.
Variant	An assessment product within the Aptis test system which shares the common framework for development and branding of other Aptis assessment products, but is treated for registration, scheduling, and scoring of candidates as a assessment product. Within the Aptis test system, the standard assessment product is Aptis General. Variants have been developed at different levels of the localisation framework, e.g. Aptis for Teachers and Aptis for Teens.
Version	Each complete, separate test form for a component within an assessment product that is considered a complete form of that component for administration to candidates, and is thus interchangeable with other complete forms of the same component. All versions of the same component of Aptis General have the same format, number of items, and types of tasks, and are constructed to have the same level of difficulty. These versions are thus considered interchangeable for any candidate taking that component of Aptis General. (In the general testing literature, what is here referred to as a *version* is often called an *alternate form* of the same test.)

BRITISH COUNCIL
APTIS TECHNICAL REPORTS

Aptis General Technical Manual
Version 1.0

Barry O'Sullivan, British Council
Jamie Dunlea, British Council

www.britishcouncil.org/aptis

ISSN 2057-7168