BRITISH COUNCIL

ENGLISH LANGUAGE
ASSESSMENT RESEARCH GROUP

Technical Report

Aptis General Technical Manual

Version 2.2

TR/2020/001

Barry O'Sullivan, British Council
Jamie Dunlea, British Council
Richard Spiby, British Council
Carolyn Westbrook, British Council
Karen Dunn, British Council

ISSN 2057-7168

www.britishcouncil.org/aptis

CONTENTS

LIST OF TABLES

LIST OF FIGURES

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 five chapters. Chapter 1 is an introduction while Chapter 2 provides an overview of the Aptis test system. Chapter 3 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. Chapter 4 describes Aptis General, divided into five subsections: Section 4.1 describes the Aptis General Revision Project; Section 4.2 gives information on the test users; Section 4.3 describes the test purpose, test structure and content, and test administration; Section 4.4 explains the scoring procedures and Section 4.5 outlines the alignment with the Common European Framework of Reference (CEFR). Chapter 5 provides an overview of the other current variants of Aptis: Aptis Advanced, Aptis for Teachers and Aptis for Teens.

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 (2020).

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 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

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

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 was 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 primarily for business-to-business (B2B) language assessment solutions. Since its inception, variants within the Aptis system have been taken globally and in some situations demand has led to direct administration to individual test takers.

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 (2011a, 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 and Weir (2009) model for reading, the model suggested by Field (2019) 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 & 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 (2011a), 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 (2020).

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 (2011a), 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 Chalhoub-Deville and O'Sullivan (2020).

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.
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 Advanced, 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 TEST PRODUCTION AND RESEARCH

3.1 Description of the test production process

3.1.1 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 4.3.4, 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 3 to 7) 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.

3.1.2 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.

3.1.2.1 The commissioning and quality review process

Only trained item writers are commissioned to write the content, which is constructed according to detailed task specifications (see Appendices B-F). Item writers have access to the test specifications on a secure online content management system, which also includes example items and templates for new items. The item writers submit a first draft of their items via a secure online file sharing platform. These items are reviewed by trained Quality Reviewers using a number code system against a set of moderation sheets derived from the specifications. The coding system, supplemented with comments from the reviewer, identifies any element of the item that does not meet any part of the specifications. Annotated items with completed moderation sheets are returned to item writers via the file sharing platform. The item writers revise their items in line with the coded feedback and comments, and resubmit the items as a second draft. The second draft submissions are reviewed by the Quality Reviewers to confirm that feedback has been acted upon appropriately. Items that pass this second quality review stage are reviewed by the Quality Assurance Managers, before being signed off by the Test Production Manager. These items are then added to the computer-based authoring system used for the creation and storage of all Aptis test tasks. 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.

3.1.2.2 The pre-testing process

All items from receptive skills components are subject to pre-testing before final availability for use in live tests. As with many large-scale standardised tests, quality assurance for productive skills takes a different approach to the receptive skills due to logistical and security issues and is maintained both through rigorous task specification at the item-writing stage and also through comprehensive rater training and standardisation.

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 Quality Assurance Manager 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 4.4.2.1 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.

3.1.2.3 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 4.4.2.1, 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 test takers are comparable across versions. Once test versions for each of the skills are constructed from items and tasks that have passed all previous stages of the test production and quality assurance cycle, they are then proof-read. As all items are constructed within a computer-delivery platform system, a final step is a full quality assurance to ensure that all system settings were accurate.

3.1.2.4 Item Writer and Quality Reviewer training and recruitment

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.

The current pool of item writers includes a number who went through the original training programme, and have amassed several years' experience of writing Aptis test items. Item writer recruitment is no longer limited to British Council staff, and applications are considered from external candidates.

The current model for item writer training is a five-week online course, which is moderated by the Quality Assurance Managers. Participants typically devote between 5 to 7 hours to each week's module. The course includes a foundation background in language testing theory and instruction followed by hands-on item writing, with a combination of peer and moderator review. There are no specific qualifications required of prospective item writers. However, all potential applicants need to demonstrate sufficient expertise in language teaching and assessment – e.g. a teaching certificate and relevant practical experience. Following the training, item writers produce example test items during a probationary period. These items are quality reviewed, and item writers are given feedback. Item writers who successfully complete the probationary period are invited to become contracted item writers.

Online training and standardisation is also provided for Quality Reviewers, who are recruited from the pool of item writers. The training is a six-week online course, which is moderated by the Quality Assurance Managers.

Regardless of the mode of delivery of the training, the core elements are standardised to participants with comprehensive training in key concepts in testing important for the process of item writing and reviewing, familiarisation with the CEFR and the test and task specifications for Aptis, as well as providing hands-on practice at item writing and reviewing. Lessons learned from the ongoing quality review process in the test production cycle have fed back into training, which has evolved over time and will continue to do so.

3.2 Accommodations

As described in Section 4.3.1, Aptis General is offered directly to organisations who wish to use it to test their employees, students, etc. 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. Test accessibility is enhanced through CBT, for example, display size and colour and audio volume can be adjusted by the test taker. Certain accommodations, if deemed appropriate, can be undertaken from 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 test takers (e.g., pen and paper over CBT)
- Braille and screen-reader compatible test versions
- amanuensis for test-takers requiring assistance with keyboard use
- extra time for test takers when this is deemed appropriate
- adapted marking procedures when criteria introduce construct irrelevant variance

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

3.3 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. These include the following reports: *Aptis Scoring System* (Dunn, 2019), *Aptis Test Development Approach* (O'Sullivan, 2015a), *Aptis Formal Trials Feedback Report* (O'Sullivan, 2015b), *Linking the Aptis Reporting Scales to the CEFR* (O'Sullivan, 2015c), *Aptis for Teens: Analysis of Pilot Test Data* (Zheng and Berry, 2015), *Aptis Technical Update 2015-2016* (British Council Assessment Research Group, 2016) and *Speaking and Writing Rating Scales Revision* (Fairbairn and Dunlea, 2017). The second strand is published as a series of Research Reports. There are currently over 20 reports published on the website covering topics including extended time limits for L2 learners, the constructs of and cognitive processes engaged by the Aptis Writing test, complexity, accuracy and fluency in the Aptis Speaking test, L1 and listening proficiency in Paired Speaking tests, using eye-tracking for the Aptis Listening test, the effects of single or double play in the Aptis Listening test, interacting with visuals in L2 Listening tests, cognitive processes involved in the Aptis Reading tests and validating the Core Inventory for General English. Both series of reports are freely available online, along with the most recent information regarding proposals which have been accepted under the ARAGs scheme and major research projects being undertaken by the Assessment Research Group, in 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 4.4.2.1, 4.4.3.5 and 3.1.2.

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: https://www.britishcouncil.org/exam/aptis/research/assessment-advisory-board.

4. 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.

4.1 Aptis General revision project

4.1.1 Background to the revision project

Aptis General was launched in 2012 and rapidly became a large-scale, international standardized test of English proficiency. From the outset, the British Council has been committed to carrying out ongoing research and to being responsive to local needs. This has led to an ongoing research agenda both internally and also through the Assessment Research Awards and Grants (ARAGs) scheme.

From the beginning, Aptis was also developed to be a dynamic system that would evolve and change as required. Thus, in 2015, a project was launched to revise the reading and listening components of Aptis General to take account of the body of research that had been collected on these two skills along with accumulated feedback from test users and global exams teams. The revision project was designed to introduce the first round of major revision changes the operational release of Aptis. As such, only changes which were practically realisable within these constraints were to be in scope. Since Aptis for Teens and Aptis Advanced were the most recently introduced variants, both of these tests benefited from a development agenda which included taking account of lessons learned from the introduction of Aptis General and Teachers. As such, it was decided to focus on Aptis General, the standard variant in the system, and Aptis for Teachers, with reading and listening identified as the highest priorities for change in this revision project. The revised Aptis General test is currently live, while the revised format of Aptis for Teachers is due to be rolled out in 2021.

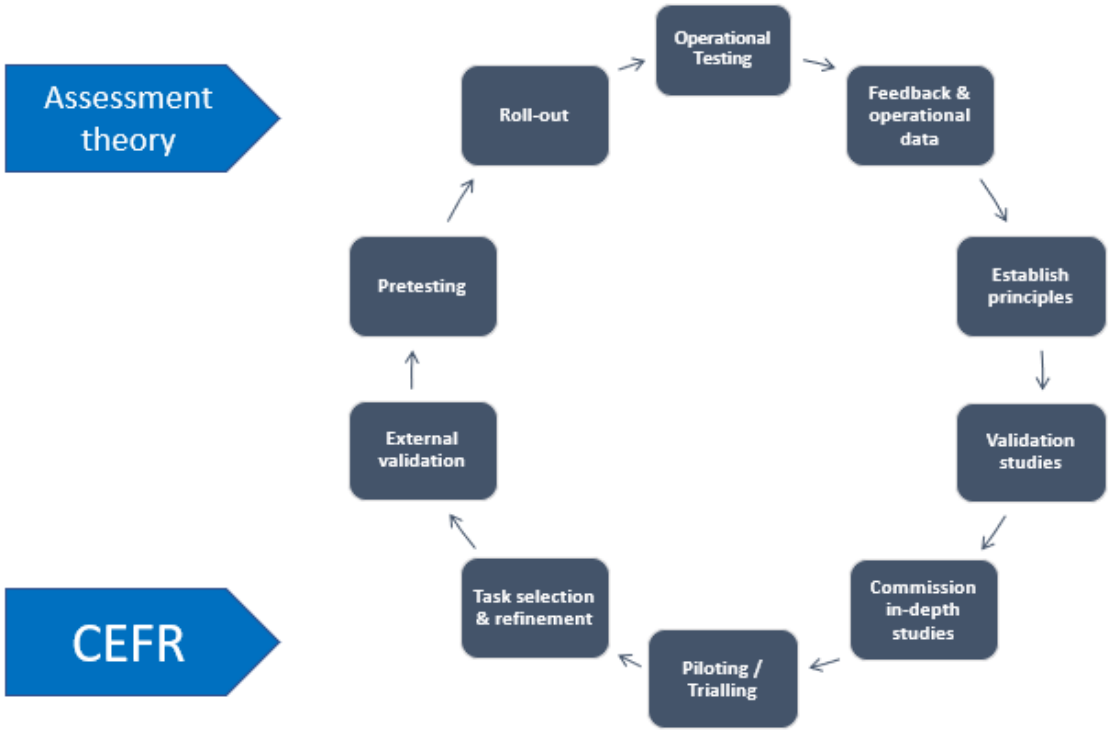
4.1.2 The revision process

The Aptis revision project consisted of multiple iterations of design, trialling and analysis conducted through collaboration between the British Council Assessment Research Group and the Assessment Development Team of British Council Global Assessments. Figure 1 provides a simplified schematic overview of the revision process, including only major activities. A more detailed explanation of activities will be published separately in a full technical report as part of the Aptis technical report series.

In keeping with the principles above, that Aptis is designed to be flexible, dynamic and evolve over time, test revision is viewed as a cyclical rather than linear process, with test research and development a constant feature of the Aptis system. As with the initial development of Aptis (O'Sullivan, 2015), the process was underpinned by advances in assessment theory, both in terms of theoretical frameworks, for example, the socio-cognitive framework of language test development and validation, and advances arising through more empirically based research. In addition, the revision was informed by the performance standards of the CEFR, both *a priori* in task development and *a posteriori* in standard setting (described in Section 4.5). In terms of the key stages in the process, first, an internal review of the test was conducted on the basis of operational data, feedback from test-takers and feedback from British Council test administrators as part of ongoing quality assurance. Once a decision was reached to move ahead with the revision process, key principles were established to identify key strengths to be preserved and prioritise areas for innovation. After this initial planning stage, validation studies were once again reviewed (e.g. Brunfaut and McCray, 2015; Holzknecht et al, 2017) to highlight specific recommendations for change. Where further insight was considered necessary, additional in-depth studies were commissioned (e.g. Field, 2015) to answer specific research questions and generate potential new item types. At this point, a selection of tasks were piloted on a small scale, then analysed, selected and refined in order to undergo larger scale field trialling, from which a greater understanding could be gained of the measurement properties of the proposed task types. After finalisation of the test design and task specifications, these together with exemplar items were submitted for external validation to prominent experts in the field of

language testing and assessment, including the Assessment Advisory Board, a panel of expert advisors appointed to periodically evaluate the research and validation activities of the Assessment Research Group. When the external review process was completed, full-scale item production was initiated, and extensive pretesting carried out in preparation for the construction of live test forms. The revised Aptis General Format was rolled out operationally in April 2020, with a revised format of Aptis for Teachers due to go live in 2021.

Figure 1: Overview of the Aptis General revision process



In line with key principles above, many of the distinctive features of the Aptis test have been retained, e.g. tasks targeted at specific CEFR levels, optional double-play in listening, The main revisions to the listening component are outlined below. See tables 3-7 for an overview of the content of the test.

Table 2: Main changes to the revised Aptis General receptive components

Component	Change	Main purpose
Listening	3 options for multiple-choice items.	To reduce ordering effects of options on candidate response behaviour
	Presentation of answer options presented in written form only.	To prevent interference of verbal presentation of options with candidate reading behaviour
	New task type targeting B1 level – matching short monologues to speakers	To broaden construct representation and introduce variety in task type at B1 level
	New task type targeting B2 level – an extended monologue inferencing meaning	To achieve greater construct representation over extended discourse at B2 level
	New task type targeting B2 level – an extended dialogue to identify speakers' opinions	To achieve greater construct representation across utterances and discourse at B2 level
Reading	Pairwise scoring model for A2 tasks, where marks are awarded for correct adjacent sentences.	To increase consistency of scoring model with A2 construct of intersentential cohesion
	A2 6-sentence reordering task replaced by two 5-sentence tasks.	To increase reliability of scores at A2 level using shorter tasks
	A new task type is added targeting B1 level – matching statements of opinion to short texts	To broaden construct representation

The information in the remaining parts of this section refer to the revised Aptis General format and so care should be taken to cross-reference between this revised technical manual 2.1 and the previous document 1.0 published online (O'Sullivan and Dunlea, 2015) to ensure there is no confusion.

4.2 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.

4.3 Test system

4.3.1 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. 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 the initial 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. These sources of information were also integrated into the test revision process.

4.3.2 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.

[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).

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.

4.3.3 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 F. 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 3: Overview of the structure of the Aptis General Core component

Part	Skill focus	Items / part	Lvl	Items/ level	Task focus	Task description	Response format
1	**Grammar**	**25**	**A1**	5	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			
			B1	5-7			
			B2	5-7			
2	**Vocabulary**	**25**	**A1**	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	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	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.
				5	Definition (vocabulary breadth)	Matching words to definitions.	5 definitions. Select the word defined from a bank of 10 options.
			B2	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 4: Overview of the structure of the Aptis General Reading component

Skill focus	Items	Marks	Lvl	Task focus	Task description	Response format
Sentence level meaning	5	5	**A1**	Sentence level meaning (Careful, local reading)	Gap fills. 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.
Inter-sentence cohesion	5	6	**A2**	Inter-sentence cohesion (Careful global reading)	Reorder 5 jumbled sentences to form a cohesive text.	Reorder 5 jumbled sentences in a 6-sentence text (the first sentence is fixed)
	5			Inter-sentence cohesion (Careful global reading)	Reorder 5 jumbled sentences to form a cohesive text.	Reorder 5 jumbled sentences in a 6-sentence text (the first sentence is fixed)
Text-level comprehension of short texts	7	7	**B1**	Text-level comprehension of short texts (Global reading, both careful and expeditious)	Matching statements of opinion with people associated with different texts. Selecting the correct person requires text-level comprehension and reading across multiple sentences.	4 short paragraphs. Test takers choose from a drop-down menu which of the four people match 7 statements.
Text-level comprehension of long text	7	7	**B2**	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 5: Overview of the structure of the Aptis General Listening component

Skill focus	Items	Lvl	Format	Task description	Response format
Lexical recognition	5	**A1**	Monologues	Q&A about listening text. Listen to short monologues (recorded messages) to identify specific pieces of information (numbers, names, places, times, etc.).	3-option multiple choice. Only the target is mentioned in the text.
Identifying specific, factual information	5	**A2**	Monologues & Dialogues	Q&A about listening text. Listen to short monologues and conversations to identify specific pieces of information (numbers, names, places, times, etc.).	3-option multiple choice. Lexical overlap between distractors and words in the input text.
Identifying specific factual information	3	**B1**	Dialogues	Q&A about listening text. Listen to short conversations to identify propositions. The information targeted is concrete and of a factual/literal nature. Requires text-level comprehension and listening across sentences/ utterances in order to answer items correctly.	3-option multiple choice. Distractors should have some overlap with information and ideas in the text. Target and distractors (where possible) are paraphrased.
	4		Monologues	Identifying aspect of a topic and matching this to a speaker. Listen to a short description to identify propositions. The information targeted is concrete and of a factual/literal nature. Requires text-level comprehension and listening across sentences/ utterances in order to answer items correctly.	Multiple matching – 4 speakers are matched with the best option from 6 written options.
Meaning representation / inference	4	**B2**	Dialogues	Matching the views of two speakers with written views on a topic. Listen to a dialogue to identify which speaker holds each attitude, opinion or intention. The information targeted should be of a more abstract nature and will require the integration of propositions across the input text to identify the correct answer.	4 items (written statements), 3 options for each: 'man', 'woman', 'both'. Targets and distractors are paraphrased, and distractors refer to important topic-related information and concepts in the text that are not possible answers to the question.
	4		Monologues	Q&As about listening text. Listen to a short talk and answer 2 questions related to the speaker's attitude, opinion or intention. The information targeted will require integration of propositions across different sections of the input text to identify correct answers.	2 x 3-option multiple choice. Both target and distractors are paraphrased, and distractors refer to information and concepts in the text that are not possible answers to the question.

Table 6: Overview of the structure of the Aptis General Speaking component

Part	Skill focus	Lvl	Task description	Channel of input / prompts	Time to plan	Time for response	Rating criteria
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 7: Overview of the structure of the Aptis General Writing component

Part	Skill focus	Lvl	Task description	Channel of input / prompts	Expected output	Rating criteria
1	**Writing at the word or phrase level. Information to simple questions in a text message type genre.**	**A1**	The candidate answers 5 simple questions. Each of the 5 responses are at the word or phrase-level.	Written. 5 short questions with space for inputting short answer responses by the candidate.	5 short gaps which can be filled by 1–5 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 fills in information on a form. 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 part 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 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	

4.3.4 Mode of delivery

Aptis General is usually taken as a computer-based test (CBT) available on PCs and tablets. 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 4.4.2.1 and 4.4.3.5 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 test takers to enhance security. The system is designed to prevent the same live version of a component being presented to the same test taker twice when the same test taker (registered once with the same details) is scheduled to take the test more than once.

4.3.5 Administration and security

Aptis General is generally sold directly to organisations, rather than 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 provides guidance and technical support for test administration. 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.

4.4 Scoring

4.4.1 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 comprehensive overview of the mechanisms and technicalities of the Aptis scoring system can be found in Dunn (2019).

4.4.1.1 Relationship between score elements

For each of the four skill components, Listening, Reading, Speaking, and Writing, a numerical scale score (between 0 and 50) plus a CEFR level are allocated to each test taker. For the Core component a numerical scale score is provided.

As noted in Section 4.3.1, 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). The cut-off scores for CEFR level designations have been set separately on the scale for each skill component; scale scores should therefore not be compared directly across skills, i.e., 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.

Table 7 shows the levels of the CEFR with the accompanying designation used for reporting in Aptis General. The level description column employs CEFR terminology to describe learner levels. 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 test taker 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 test taker who receives a near perfect score will receive a level classification of C. This means the test taker 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; this cut off was explicitly addressed in the standard setting exercise for each of the four skill components (O'Sullivan, 2015b). Aptis General does not distinguish between C1 and C2. For test takers requiring discrimination at these levels, other Aptis variants, e.g., Aptis Advanced, are available.

Table 8: 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

Note that a CEFR level is not reported for the Grammar and Vocabulary component. The Core component assesses test takers' grammar and vocabulary knowledge. Since this knowledge underpins all language skills (see McCray & Dunn, 2020), it is an essential component in the Aptis testing system. However, CEFR levels are not reported for the Core component at the current time, because the position of grammar and vocabulary knowledge within the CEFR is one of the most under-specified elements of the framework. The Core component does nonetheless play a role in the CEFR level allocation system for each skill component, as is elaborated at length in 4.4.5 below. The

Core component is therefore an essential element in all packages of the Aptis test, and CEFR level allocation will not be finalised for any test takers who do not complete this component.

4.4.1.2 Score reporting and interpretation

For each of the four skill components, a skills profile is provided to test takers which reports both a numerical scale score (between 0 and 50) and a CEFR level. These pieces of information are useful for different purposes, as summarised below.

Numerical Score:

- Provides a detailed comparison of test-taker performances for a given skill within a group, including comparisons between students within the same CEFR level at a more fine-grained level.
- Enables tracking of test taker performance for a given skill over a period or following language teaching/learning intervention. This is particularly relevant when the intervention or learning period may not be sufficient to realise improvement over one or more CEFR levels.

CEFR Skill Profile:

- Provides benchmarked CEFR levels of proficiency which can be referenced to descriptions of what a language user can typically do at these levels.
- Differentiates strengths and weaknesses across skills to help provide road maps for learners and teachers to target areas for improvement (referencing the descriptions of what typical language users can do).
- Can be used to show improvement over longer periods of time or more intensive interventions based on recognized criteria.

Test takers who complete the four-skills test, comprising all five Aptis components, are additionally awarded an overall numerical scale score (out of a total possible score of 200) and an overall CEFR level.

4.4.2 Reliability of receptive skill components

Two key indicators commonly reported for testing programmes are the reliability and the Standard Error of Measurement (SEM). 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 (*AERA, APA, 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. *AERA, APA, 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 4.4.3.4.

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).

A useful measure for interpreting the accuracy of individual scores is the SEM. 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. 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. The smaller the number for the SEM, the more accurate the test will be.

Estimates of reliability and SEM have been calculated for test versions using the revised format of the Aptis test. To derive these estimates, reliability and SEM were calculated for multiple test forms which were used in operational testing. The average estimates for these test forms are shown in Table 9. 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 therefore demonstrate appropriate levels of reliability and SEM. It should be remembered that these estimates were derived from an initial subsample of global data from the first months of the live testing program following the launch of the revised listening and reading components. These figures will be further updated when a larger sample of operational data becomes available.

Table 9: Mean reliability and SEM estimates for pre-testing versions of Reading and Listening tasks

	Listening	Reading
Mean	0.83	0.86
SEM	3.83	4.03

4.4.2.1 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 3.1.2 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.

4.4.3 Reliability of productive skill components

4.4.3.1 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. A team of Assistant Examiner Managers and Senior Examiners work under the guidance of the Examiner Network Manager to 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.

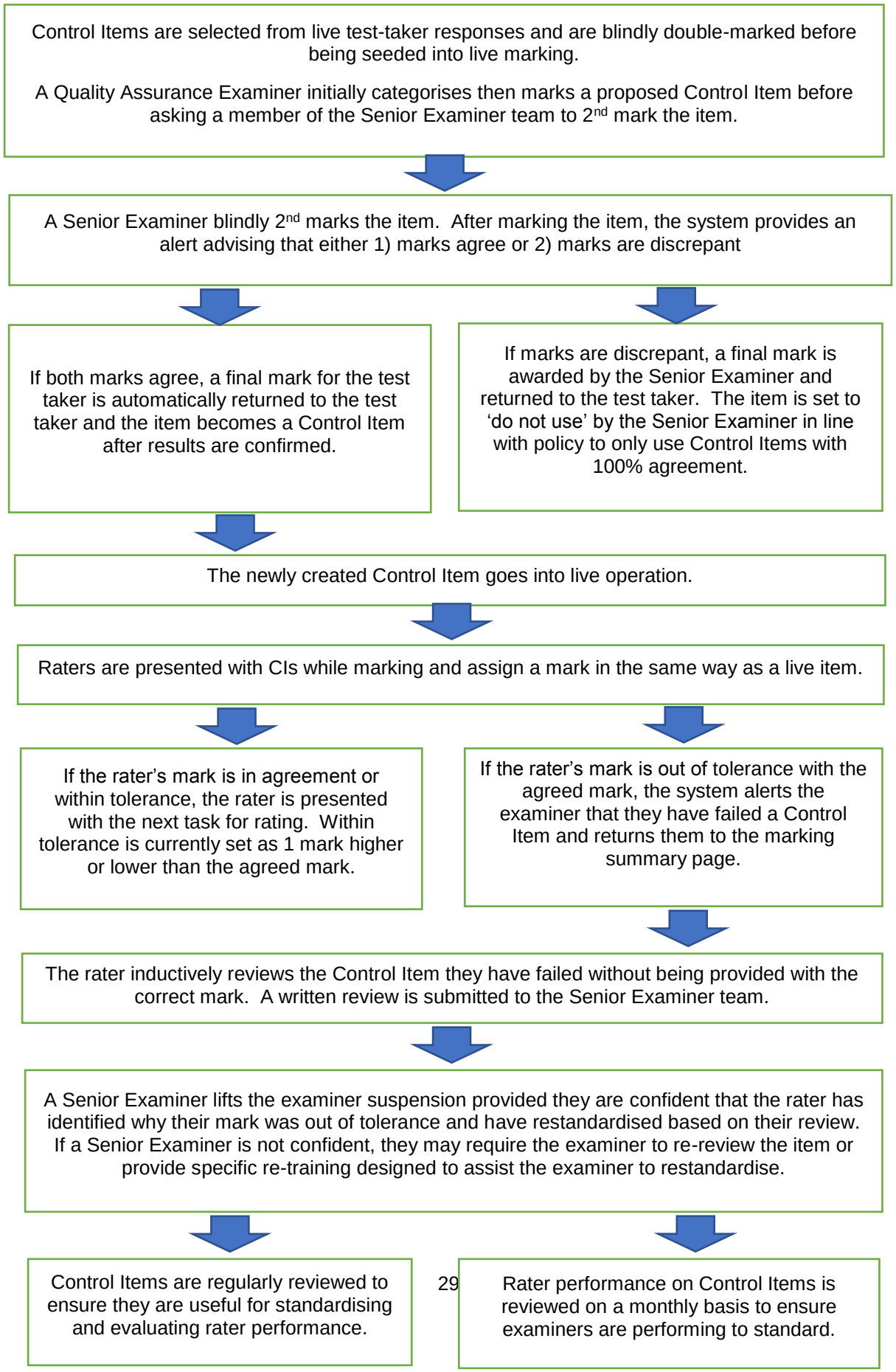
The online rating system automatically breaks up a test-taker's performance on a full Speaking or Writing test into the separate responses for each task (see Table 6 and Table 7 for an overview of the tasks in each component). The same rater will not be able to rate more than one task performance for 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 2 has been adapted from Fairbairn (2015) to provide an overview of how the CI system works in practice.

Figure 2: Overview of control item (CI) system (from Fairbairn, 2015, revised by Catherine Hughes in Feb 2020)



4.4.3.2 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 (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.

4.4.3.3 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).

Writing Task 1 is marked on a scale of 0-3. Writing Tasks 2 and 3 and Speaking Tasks 1-3 are marked on a scale of 0-5. 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).

4.4.3.4 Inter-rater reliability

As outlined in Section 4.4.3.1 above, 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 an individual 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 4.4.2 above, 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. As such, the inter-rater correlations generated by this study were thus likely higher than the correlations that would be seen for ratings based on a more varied sample of performances, which include more borderline and problematic examples. Nevertheless, while this study had important limitations, the use of CI data to investigate inter-rater reliability represents an innovative way to obtain rating data from multiple raters on the same items under operational rating conditions.

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

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 outlined above thus offered a way to gain insights into rater consistency under operational conditions. However, a clear need to follow up with further studies was recognised, including specially designed multiple-rating studies which would necessarily be carried out in an experimental setting outside the normal operational rating environment.

Subsequently, Fairbairn and Dunlea (2017) carried out such a multiple rater study as part of field trialling for the Aptis Speaking and Writing Rating Scales Revision project. This study aimed to use rater feedback from the rating process during the first year of operational test use to inform revisions to the rating scales, and to validate the new scales through piloting and field trialling. The project focused on 'improving the clarity and usability of the rating scales' (p. 12) and involved a cyclical approach to the process of scale revision, including collecting data from all operational raters via a questionnaire, followed by a focus group with assessment experts. Once new rating scales had been developed, a small-scale pilot was carried out involving seven senior raters marking 12 writing and speaking samples. Data was then collected from five raters via a focus group and the rating scales were fine-tuned accordingly. After this, a field trial of the new rating scales was carried out, involving 49 raters marking 100 writing scripts and 30 speaking responses across CEFR levels A1-C.

The data collected from the field trial was analysed using multi-faceted Rasch measurement (MFRM) analysis to investigate rating quality. MFRM provides measures of fit to the Rasch model, which are indicators of consistency of rating among all raters in the sample. Raters can be classified as exhibiting good fit or as misfitting according to the infit mean square value ascribed to them in the analysis. Values between 0.6 and 1.5 are considered to be acceptable, while values outside this range

denote misfit (Eckes, 2011; Lunz, Wright and Linacre, 1990). Table 11 shows the number of misfitting raters for the writing and speaking field trials (see Fairbairn and Dunlea, 2017 for a more detailed description of the results).

Table 11: Summary of rater fit for Writing and Speaking field trial

Component	Number of raters	Number of misfitting raters
Writing	49	1
Speaking	49	3

As can be seen from the table, in rating the writing papers, only one rater was identified as misfitting. This indicates that raters exhibited consistent behavior using the scale. For the speaking tasks, there were three raters outside of the acceptable parameters for infit mean square, which again demonstrates that the majority of raters were using the revised rating scales consistently.

Relative rater severity could also be assessed using MFRM in the field trial, as measures on the logit scale provided for each rater denote their leniency or harshness in comparison to the other raters in the sample. Results showed that for both speaking and writing, raters were clustered around the mean, with the vast majority of raters within an acceptable range of +/-1 logit. For writing only one rater and for speaking five raters were outside this range, the latter finding being attributed to the greater extent of revisions made to the speaking scales. Despite a small degree of rater variation for speaking, the results were taken to indicate that both scales could be used consistently.

This large-scale experimental project therefore further supports a high level of rater consistency for the Aptis General speaking and writing tests, and the study has provided useful validation evidence for the revised rating scales.

4.4.3.5 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 out 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.

4.4.4 Precision of scoring: Standard Error of Measurement

As noted in Section 4.4.2, 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. 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 12 provides estimates of the average SEM for operational versions for each of the five components of Aptis General.[4]

Table 12: Estimates of Standard Error of Measurement (SEM) for Aptis General component

	Core G&V	Listening	Reading	Speaking	Writing
Scale score	0–50	0–50	0–50	0–50	0–50
SEM	2.97	3.83	4.03	3.7	2.0

4.4.5 CEFR level allocations

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 two standard-setting studies 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 first standard-setting study are reported in a separate technical report (O'Sullivan, 2015b). The second study, concerned with standard-setting for the revised listening and reading components, will be published in 2020.

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.

4.4.5.1 The role of the Core component in CEFR level allocation

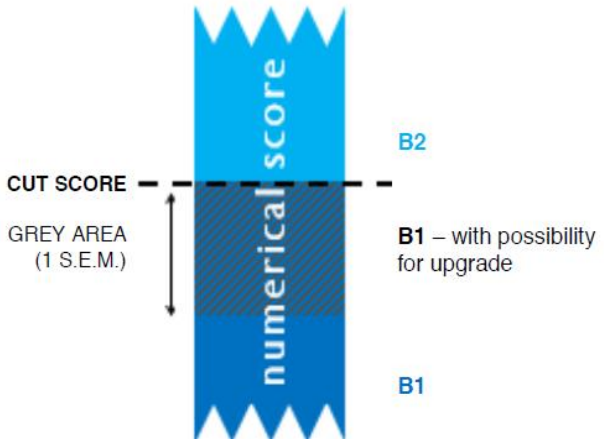
In cases in which a test taker's performance in any of the four skill areas falls just shy of a grade boundary, score information from this Core component is used to determine whether a given test taker should remain at the lower CEFR level or be upgraded. This procedure is intended to increase the fairness and accuracy of grade allocation, and reflects the understanding of grammar and vocabulary as key sub-processes in models of L2 language ability (Field, 2013; Khalifa & Weir, 2009). In this respect, performance on the Core component is associated with some of the fundamental skills required in each of the skill components, and the use of the score information to refine decisions is justified on theoretical grounds (O'Sullivan & Dunlea, 2015b). Empirical investigations have been conducted using a large global Aptis dataset in order to explore the functioning of this procedure; this

[4] 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 Cronbach Alpha 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 10 were used as the reliability estimates.

project is described in detail in McCray and Dunn (2020). This study mapped the relationship between grammar and vocabulary, and each of the skills of listening, reading, writing, and speaking for a wide range of abilities spanning CEFR levels A0 to C, and showed the relationship between grammar and vocabulary to hold across the ability spectrum.

The Core language knowledge component score is drawn upon in CEFR level allocation when a test taker achieves a score on one of the main skills components that falls within one standard error of measurement (SEM) of a CEFR level boundary. The score on the Core component will determine whether the test taker will remain at the lower CEFR level or whether they will be upgraded to the higher level. To receive this upgrade, they should perform significantly above the average on the Core component (set as one standard deviation above the mean). The process is illustrated in Figure 3 below.

Figure 3: Illustration of the "grey area" in which candidate CEFR level allocation is contingent on Core component performance



.

This review and adjustment is undertaken automatically within the system. It is important to note that this process does not affect the reported scores on the scale of 0–50 in the relevant skills component for test takers. It is therefore possible for two test takers to receive the same numerical score for a skills component, and a different CEFR allocation. This will only be the case if both test takers achieve a score close to the cut score between two CEFR levels, but one of the test takers performs significantly better in the Core component. Please refer to Dunn (2019) for further discussion on Aptis scoring mechanisms.

4.4.5.2 Overall CEFR level allocation

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 test taker's proficiency profile, in addition to the overall CEFR level.

4.5 Standard setting and linking to the CEFR

The following sections provide an overview of the second standard-setting study, conducted in order to determine cutscores for the revised listening and reading components of Aptis General.

4.5.1 The role of standard setting

Alignment of an assessment with standards for reporting and policy implementation is a comprehensive activity involving both qualitative and quantitative procedures. Qualitative data collection begins with a comprehensive evaluation of the content relevance and alignment of the knowledge and competencies measured by a test with the way those same features are described in the set of standards which is the target of alignment. In this project, the set of standards in focus is the CEFR. The central quantitative data collection in the alignment process is the process of standard setting (Cizek & Bunch, 2007; Council of Europe, 2009; Dunlea et al, 2019). Standard setting has its origins in the educational measurement tradition in the United States, but has also been widely applied, and adapted, in the field of language testing due to the rapid spread of the CEFR (Council of Europe, 2001) in education systems internationally.

4.5.2 Overview of the alignment procedure

The methodology used to align the revised reading and listening components of Aptis General with the CEFR drew on an extensive body of literature from this field. The study benefited from the ability to draw on the theoretical expertise and direct operational experience of the British Council's Assessment Research Group in the first study linking Aptis to the CEFR (O'Sullivan, 2015b; O'Sullivan and Dunlea, 2015), as well in other large-scale linking and test comparability projects involving locally developed national standards in Asia (Dunlea et al., 2018; Dunlea et al., 2019).

As noted above, standard setting is at the core of the linking process. However, the process of linking encompasses more than standard setting alone. The Council of Europe's *Manual for Linking Exams to the CEFR* (2009) specifies five stages through familiarisation, specification, standardisation, standard setting and validation. The theoretical framework developed by the British Council draws on work carried out in a range of contexts internationally by Dunlea (2015), Dunlea and Figueras (2012), Dunlea et al. (2019), and O'Sullivan (2015b), to synthesise these steps into three main evidence-collection categories:

- construct definition: gathering evidence of the alignment of the constructs underpinning the test and target standards
- standard setting: gathering empirical data to drive the statistical basis for setting cutoffs of the test score scale which represent criterial levels within the target standards
- validation: synthesis of internal and external evidence to support the standard-setting process.

This three-stage theoretical framework for linking exams to standards is described more fully in Dunlea et al. (2019). The process of aligning the revised reading and listening components of Aptis General with the CEFR covered these three key categories, with a principal focus on the second, standard setting.

Construct definition involves first making detailed evaluations of the test at the task and item level. Test tasks and items are evaluated using a comprehensive set of criterial features refined from the Aptis test specifications, similar to the test analysis grids in the *Manual for Linking Exams to the CEFR* (Council of Europe, 2009), This detailed evaluation of the skills and abilities targeted by the items provides an explicit set of features which can then be compared to the description of proficiency included in the CEFR level descriptions. Thus, in this study, trained groups of expert judges with experience in language test development and research evaluated each task and item of a complete test for each of the receptive skills components (reading, listening) of Aptis General. They identified specific CEFR Performance Level Descriptors which they judged were operationalized by these tasks and items, providing a theoretical basis for the subsequent stages of the linking process.

The standard-setting methods used in the study have been documented extensively in relation to linking exams to standards. For receptive skills components, two test-centred methods were employed, in which a panel of expert judges identified the level of test-taker attainment required at

each performance standard, ie. CEFR level. These were the Basket method and the Modified Angoff method, both of which are frequently used and have been widely researched in relation to the CEFR (O'Sullivan, 2015b). The two methods were used in combination, as in the approach developed by Dunlea (2015). Initial Basket method judgements served as reference points to help judges make more refined decisions using the Modified Angoff method. These decisions were then analysed using Multi-Faceted Rasch Measurement (MFRM), and the results were used to determine final cutscore recommendations.

Validation in the context of alignment is concerned with gathering evidence to support the validity of the alignment process rather than with validation of the test itself. Accordingly, three sources of evidence were used to validate the linking clams and final cutoff estimates for Aptis. Firstly, procedural validity evidence was gathered from descriptions of methodological processes, training procedures and questionnaire feedback from participants, indicating that robust processes had been understood and implemented by panelists. Then, internal validity evidence was provided through MFRM analysis of the consistency and accuracy of the results, identifying that participants converged toward a common standard over the course of standard-setting rounds of judgements. Finally, the external validity of the process was supported by comparing these results with others obtained from other standard-setting methods and frameworks, such as China's Standards of English (Dunlea et al., 2019)

A full description of the methodology, data collection, analysis and results of the project to align the revised components of Aptis General to the CEFR is provided in a separate technical report to be published in 2020.

4.5.3 CEFR alignment results for Aptis General

According to the procedures outlined in section 4.5.2 above, the resulting cutscores used in scoring for Aptis General on the Common European Framework of Reference are presented below. The cutscores in Table 13 represent the starting point of each level on the 0–50 scale for each test.

Table 13: CEFR cutscores for Aptis General

	A1	A2	B1	B2	C
Listening	8	16	24	34	42
Reading	8	16	26	38	46
Writing	6	18	26	40	48
Speaking	4	16	26	41	48

Cutscores for the other variants of the Aptis test system can be found in *Aptis Scoring System* (Dunn, 2019).

5. OVERVIEW OF OTHER APTIS VARIANTS

5.1 Aptis Advanced

Aptis Advanced is designed to provide assessment options for ESL/EFL speakers spanning proficiency ranges from B1 to C2 in terms of the Common European Framework of Reference for Languages (CEFR). Test-takers will be 16 years old or older and may be engaged in education, training, employment or other activities.

As with Aptis General, the description of test-taker variables for Aptis Advanced is generic. It is intended as a ready-to-use product (levels 0–1 of the localisation framework), appropriate for use in a broad range of contexts. Potential test users are expected to engage with the Aptis team to evaluate whether Aptis Advanced is the most appropriate variant for the intended test-taker population.

Aptis Advanced is intended for use in determining the ability of test-takers at higher proficiency levels (B1-C2) in a range of employment, training and learning needs. Potential target language use (TLU) contexts lie within the educational, occupational, and public domains, for example, where learners are engaged with real-world tasks in higher education and training programmes, as well as learners using English for work-related purposes. See Section 4.3.1 for typical uses for which the test may be considered appropriate.

Tables 14 to 18 present an overview of the structure of the five components which make up the full, four-skills package of Aptis Advanced:

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

The Core component is always included as a compulsory component and used in combination with the other skills as required by the test user in accordance with levels 0-1 of the localization framework (see Section 2.3).

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.

Table 14: Overview of the structure of the Aptis Advanced Core component

Part	Skill focus	Items / part	Lvl	Items/ level	Task focus	Task description	Response format
1	**Grammar**	**25**	**A1**	5	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			
			B1	5-7			
			B2	5-7			
2	**Vocabulary**	**25**	**A1**	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	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	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.
				5	Definition (vocabulary breadth)	Matching words to definitions.	5 definitions. Select the word defined from a bank of 10 options.
			B2	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 15: Overview of the structure of the Aptis Advanced Reading component

Part	Skill focus	Items	Lvl	Task focus	Task description	Response format
1	**Text-level comprehension of short texts**	**7**	**B1**	Text-level comprehension of short texts (Global reading, both careful and expeditious)	Matching statements of opinion with people associated with different texts. Selecting the correct person requires text-level comprehension and reading across multiple sentences.	4 short paragraphs. Test takers choose from a drop-down menu which of the four people match 7 statements.
2	**Text-level comprehension of long text**	**7**	**B2**	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.
3	**Text-level comprehension of a shorter text**	**5**		Text-level comprehension and cohesion (careful global reading)	Selecting the correct options to complete a cloze text. There are 5 gaps and selecting the correct option can only be deduced from a global understanding of the whole text.	5 gaps and 3 MCQ options for each. Select the correct option to fill in the gap.
4	**Text-level comprehension across two texts**	**6**	**C1**	Text-level comprehension across two texts (global reading, both careful and expeditious)	Selecting the correct option to complete two thematically linked cloze texts. Selecting the correct option requires global understanding of both texts.	3 gaps in each text with 3 MCQ-options for each. Select the correct option to fill the gap.

Table 16: Overview of the structure of the Aptis Advanced Listening component

Part	Skill focus	Item/ Part	Lvl	Format	Task description	Response format
1	**Identifying specific factual information**	**5**	**B1**	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 text-level comprehension and listening across sentences/ utterances in order to answer items correctly.	One 4-option multiple choice question. Distractors should have some overlap with information and ideas in the text. Target and distractors (where possible) are paraphrased.
2	**Meaning representation / inference**	**6**	**B2**	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.	Two 4-option multiple choice questions. 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.
3	**Discourse construction, meaning representation and inference**	**6**	**C1**	Dialogues	Q&A about listening text. Listen to a dialogue between two speakers and identify which opinions are expressed by which speaker(s). The information targeted will require the integration of abstract ideas and propositions across an extended stretch of interaction.	Identify who expresses each of the six given opinions: the male speaker, the female speaker, or both the male and female speaker.
4	**Discourse construction, meaning representation and inference**	**8**		Monologues	Q&A about listening text. Listen to a monologue in which the speaker recounts a narrative containing four key elements. The information targeted will require the integration of information and propositions across an extended stretch of interaction.	Select the appropriate response from a bank of 3 MCQ options for each of the four key story elements.

Table 17: Overview of the structure of the Aptis Advanced Speaking component

Part	Skill focus	Lvl	Task description	Channel of input / prompts	Time to plan	Time for response	Rating criteria
1	**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	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	**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	
3	**Integrating ideas regarding an abstract topic into a long turn. Giving opinions, justifying opinions, giving advantages and disadvantages.**	**C1**	The candidate plans a long turn formulating a balanced argument on a topic based on input of for/against bullet points. The candidate speaks for two minutes to present his/her long-turn. A subsequent follow-up statement related to the topic is presented to the candidate once the long term has been completed. The candidate is invited to comment on the statement and has 45 seconds for their response, for which there is no preparation time.	Written and aural input (no visuals). The title of the topic is shown on screen above two tables of three 'for' and three 'against' bullet points. The follow up statement (pre-recorded) and prompt appear on screen once the long term has been completed.	1 response of 90 seconds 1 response of 45 seconds	1 minute to prepare for first response, immediate response following second prompt	

Table 18: Overview of the structure of the Aptis Advanced Writing component

Part	Skill focus	Lvl	Task description	Channel of input / prompts	Expected output	Rating criteria
1	**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.	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	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	**Integrated writing task requiring longer paragraph level writing in response to an email and some notes provided. Appropriate use of register.**	**B2**	The candidate writes an e-mail in response to the task prompt which contains an e-mail from an unknown reader connected to the information in the prompt (management, customer services, etc.) and notes made by the e-mail writer. The candidate will be required to expand these notes into complete sentences framed in an appropriate formal register.	A transactional e-mail message is presented as the starting point. This e-mail is written in a formal impersonal register. The e-mail contains three distinct points of information. The notes that accompany the e-mail are written as bullet points and/or in note form in an informal register. There are three separate notes – one for each distinct point of information in the e-mail. Number annotations indicate which notes apply to which pieces of information. The notes appear in the same sequence as the information in the e-mail.	120-150 words	
3	**Integrated writing task requiring longer paragraph level writing in response some notes provided on a given subject. Appropriate use of register for intended audience.**	**C1**	The candidate writes an informational text for an online publication on a topic of general interest	The candidate is presented with some notes in bullet point format on the topic and a simple grid (three rows, three columns) containing additional information in numerical form. The information in the bullet point notes should focus on abstract concepts. The information in the table should focus on concrete information and should be such that it allows for contrast and comparison and interpretation.	180-220 words	

5.2 Aptis for Teachers

Aptis for Teachers 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 adults engaged in education-related training, employment or other activities.

Aptis for Teachers is designed specifically to assess the English proficiency of teachers and other test takers who are working in the education sector. It is intended as a ready-to-use product (levels 0–1 of the localisation framework), appropriate for use in a range of educational contexts for the age group specified. Potential test users are expected to engage with the Aptis team to evaluate whether Aptis for Teachers is the most appropriate variant for the intended test-taker population.

Aptis for Teachers is provided directly to Ministries of Education and educational institutions. Potential target language use (TLU) contexts lie within the educational and public domains, for example, where learners are engaged with real-world tasks in schools and universities, teacher-training programmes, and other teaching-related contexts. There are a variety of typical uses for which the test is considered appropriate:

- ensuring reliable entrance and exit requirements for higher education courses
- streaming according to proficiency level within language training and teacher-training programmes
- evaluating progress within training programmes
- identifying individuals with the language proficiency levels necessary for employment in different roles
- identifying strengths and weaknesses to inform teaching and improve training programmes

Tables 19 to 23 present an overview of the structure of the five components which make up the full, four-skills package of Aptis for Teachers:

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

The Core component is always included as a compulsory component and used in combination with the other skills as required by the test user in accordance with levels 0-1 of the localization framework (see Section 2.3).

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.

Table 19: Overview of the structure of the Aptis for Teachers Core component

Part	Skill focus	Items / part	Lvl	Items/ level	Task focus	Task description	Response format
1	**Grammar**	**25**	**A1**	5	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			
			B1	5-7			
			B2	5-7			
2	**Vocabulary**	**25**	**A1**	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	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	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.
				5	Definition (vocabulary breadth)	Matching words to definitions.	5 definitions. Select the word defined from a bank of 10 options.
			B2	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 20: Overview of the structure of the Aptis for Teachers Reading component

Part	Skill focus	Items	Lvl	Task focus	Task description	Response format
1	**Sentence level meaning**	**5**	**A1**	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. TLU domain is relevant for teachers.	3-option multiple choice for each gap.
2	**Inter-sentence cohesion**	**6**	**A2**	Inter-sentence cohesion (Careful global reading)	Re-order 6 jumbled sentences to form a cohesive text. TLU domain is relevant for teachers.	Re-order 6 jumbled sentences. All sentences must be used to complete the text.
3	**Text-level comprehension of short texts**	**7**	**B1**	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**	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. TLU domain is relevant for teachers.	7 paragraphs forming a long text. Select the most appropriate heading for each paragraph from a bank of 8 options.

Table 21: Overview of the structure of the Aptis for Teachers Listening component

Skill focus	Item/ Part	Lvl	Format	Task description	Response format
Lexical recognition	**10**	**A1**	Monologues	Q&A about listening text. Listen to short monologues (recorded messages) to identify specific pieces of information (numbers, names, places, times, etc.) TLU domain is relevant for teachers.	4-option multiple choice. Only the target is mentioned in the text.
Identifying specific, factual information	**5**	**A2**	Monologues & Dialogues	Q&A about listening text. Listen to short monologues and conversations to identify specific pieces of information (numbers, names, places, times, etc.) TLU domain is relevant for teachers.	4-option multiple choice. Lexical overlap between distractors and words in the input text.
Identifying specific factual information	**5**	**B1**	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. TLU domain is relevant for teachers.	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**	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. TLU domain is relevant for teachers.	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 22: Overview of the structure of the Aptis for Teachers Speaking component

Part	Skill focus	Lvl	Task description	Channel of input / prompts	Time to plan	Time for response	Rating criteria
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 23: Overview of the structure of the Aptis for Teachers Writing component

Part	Skill focus	Lvl	Task description	Channel of input / prompts	Expected output	Rating criteria
1	**Writing at the word or phrase level. Information to simple questions in a text message type genre.**	**A1**	The candidate answers 5 simple questions. Each of the 5 responses are at the word or phrase-level.	Written. 5 short questions with space for inputting short answer responses by the candidate.	5 short gaps which can be filled by 1–5 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 fills in information on a form. 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 part 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 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	

5.3 Aptis for Teens

Aptis for Teens 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 13-17 years old and will be in formal education in lower-secondary, middle school or junior high school, depending on geographical context.

Aptis for Teens is designed specifically to assess the English proficiency of students within secondary education. It is intended as a ready-to-use product (levels 0–1 of the localisation framework), appropriate for use in a range of educational contexts for the age group specified. Potential test users are expected to engage with the Aptis team to evaluate whether Aptis for Teens is the most appropriate variant for the intended test-taker population.

Potential target language use (TLU) contexts lie within the educational and public domains, in EFL/ESL contexts where English is studied at school and/or in language learning programmes outside school. Test-takers may be learning the language as a subject of study or as a medium of instruction to study other subjects. Typical uses for which the test is considered appropriate include:

- streaming learners into language classes according to proficiency level
- evaluating progress within learning programmes
- assessing strengths and weaknesses of learners to inform teaching and support
- assessing readiness of students to study in English-taught programmes
- assessing readiness for taking high-stakes certificated exams

Tables 24 to 28 present an overview of the structure of the five components which make up the full, four-skills package of Aptis for Teens:

6. Core Grammar and Vocabulary component
7. Listening component
8. Reading component
9. Speaking component
10. Writing component.

The Core component is always included as a compulsory component and used in combination with the other skills as required by the test user in accordance with levels 0-1 of the localization framework (see Section XX).

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.

Table 24: Overview of the structure of the Aptis for Teens 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 25: Overview of the structure of the Aptis for Teens Reading component

Skill focus	Items	Lvl	Task focus	Task description	Response format
Sentence level meaning	**5**	**A1**	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.
Inter-sentence cohesion	**6**	**A2**	Inter-sentence cohesion (Careful global reading)	Reorder 6 jumbled sentences to form a cohesive text	Reorder 6 jumbled sentences. All sentences must be used to complete the text.
Text-level comprehension of short texts	**7**	**B1**	Text-level comprehension of short texts (Careful global reading)	Matching statements of opinion with people associated with texts on different topics, e.g., travel, parental rules, school canteens, etc. Selecting the correct person requires text-level comprehension and reading across multiple sentences.	4 short paragraphs. Test takers choose from a drop-down menu which of the four people match 7 statements.
Text-level comprehension of long text	**7**	**B2**	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 26: Overview of the structure of the Aptis for Teens Listening component

Skill focus	Items	Lvl	Format	Task description	Response format
Lexical recognition	**5**	**A1**	Monologues	Q&A about listening text. Listen to short monologues (recorded messages) to identify specific pieces of information (numbers, names, places, times, etc.).	3-option multiple choice. Only the target is mentioned in the text.
Identifying specific, factual information	**7**	**A2**	Monologues & dialogues	Q&A about listening text. Listen to short monologues and conversations to identify specific pieces of information (numbers, names, places, times, etc.)	3-option multiple choice. Lexical overlap between distractors and words in the input text.
Identifying specific, factual information	**7**	**B1**	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 text-level comprehension and listening across sentences/ utterances in order to answer items correctly.	3-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	**6**	**B2**	Monologues	Q&A about listening text, with 2 questions per text. Listen to a talk/class presentation, etc. to identify problems, issues, solutions or recommendations which are expressed by the speaker. The information targeted will require integration of propositions across different sections of the input text to identify correct answers.	2 x 3-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 27: Overview of the structure of the Aptis for Teens Speaking component

Part	Skill focus	Lvl	Task description	Channel of input / prompts	Time to plan	Time for response	Rating criteria
1	**Giving personal information**	**A1/A2**	Candidate responds to three questions on personal topics. Each question is presented separately, and the candidate records his/her response before the next question is presented.	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).	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**	**A2/B1**	The candidate responds to three prompts/questions. 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 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) A single photograph of a scene related to the topic of the questions and familiar to A2/B1 candidates of the target age group is presented on screen.	No	45 seconds to respond to each question	
3	**Describing, comparing and contrasting, providing reasons and explanations**	**B1**	The candidate responds to 2 questions, contrasting and comparing two photographs on a topic familiar to B1 candidates of the target age group. The candidate must express and support an opinion/preference about a topic related to the photographs.	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 opinions, justifying opinions, advantages and disadvantages.**	**B2**	The candidate plans a long turn integrating information given to them and adding their own opinion/knowledge of the subject. The candidate speaks for two minutes.	The candidate is presented with a poster containing bulleted information points on the topic, which they are told they have prepared and must present to their class.	90 sec	2 minutes for the entire response.	

Table 28: Overview of the structure of the Aptis for Teens Writing component

Part	Skill focus	Lvl	Task description	Channel of input / prompts	Expected output	Rating criteria
1	**Writing at the word or phrase level. Information to simple questions in a text message type genre.**	**A1**	The candidate answers 5 simple questions. Each of the 5 responses are at the word or phrase-level.	Written. 5 short questions with space for inputting short answer responses by the candidate.	5 short gaps which can be filled by 1–5 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 fills in information on a form. The candidate must write short responses 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 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 purpose/ activity used in part 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	**Continuous writing task requiring essay level writing. Responding to a prompt on a topical issue.**	**B2**	The candidate writes a short essay of 220-250 words in response to the task prompt which contains a notice asking for essay competition entries. The prompt asks for an argumentative essay on a topic which Aptis for Teens test takers are likely to encounter in the public/educational domains. The topic field will be related to the same background setting used in parts 2, & 3.	Written. The instructions are presented as a short notice advertising an essay competition. The prompt will clearly identify the purpose, context, and audience of the essay competition, describe the topic and essay (task) requirements.	220-250 words. Must be in essay format with an introduction and conclusion.	

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: 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. Doi: 10.1207/s15434311laq0201_1

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

British Council Assessment Research Group. (2016). *Aptis Technical Update 2015-2016*. Retrieved from: https://www.britishcouncil.org/exam/aptis/research/publications/technical-report.

Brunfaut, T., & McCray, G. (2015). *Looking into test-takers' cognitive processes whilst completing reading tasks: a mixed-method eye-tracking and stimulated recall study*.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing 20*(4), 369–383. Doi: 10.1191/0265532203lt264oa

Chalhoub-Deville, M., & O'Sullivan, B. (2020). *Validity: Theoretical development and integrated arguments.* Sheffield: Equinox.

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., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3–13. Doi: 10.1111/j.1745-3992.2009.00165.x

Cizek, G. J., & Bunch, M. B. (2007). Standard setting. *Thousand Oaks, CA: Sage*.

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

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

Davidson, F., & 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. Doi: 10.1017/S0261444807004351

Dunlea, J., & Figueras, N. (2012). Replicating results from a CEFR test comparison project across continents. In D. Tsagari and I. Csepes (Eds.), *Collaboration in language testing and assessment* (pp. 31–45). New York: Peter Lang.

Dunlea, J. (2015). *Validating a set of Japanese EFL proficiency tests: demonstrating locally designed tests meet international standards* (Unpublished PhD thesis). University of Bedfordshire, Bedfordshire.

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

Dunlea, J., Spiby, R., Nguyen, T. N. Q., Nguyen, T. Q. Y., Nguyen, T. M. H., Nguyen, T. P. T., Thai, H.L.T., & Bui, T. S. (2018). Aptis-VSTEP Comparability Study: Investigating the usage of two EFL

tests in the context of higher education in Vietnam. *British Council Validations Series* (VS/2018/001). London: British Council.

Dunlea, J., Spiby, R. Wu, S., Zhang, J., & Cheng, M. (2019). *Technical report on the linking of UK exams to the China Standards of English*. London: British Council.

Dunn, K. (2019). Aptis scoring system. *Technical report (TR/2019/001).* London: British Council.

Eckes, T. (2011). Introduction to many-facet Rasch measurement: Analyzing and evaluating rater mediated assessments. Frankfurt, Germany: Lang.

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.

Fairbairn, J., & Dunlea, J, (2017). Speaking and writing rating scales revision. *Technical report (TR/2017/001).* London: British Council.

Field, J. (2013). Cognitive validity. In L. T. A. Geranpayeh (Ed.), *Examining listening* (pp. 77–151). Cambridge: Cambridge University Press.

Field, J. (2015). *Aptis test of listening: Final report on revision project with recommendations.* Internal British Council report: unpublished

Field, J. (2019). *Rethinking the second language listening test: From theory to practice.* British Council Monographs. *London: British Council and Equinox*.

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

Holzknecht, F., Eberharter, K., Kremmel, B., Zehentner, M., McCray, G., Konrad, E., & Spöttl, C. (2017). *Looking into listening: Using eye-tracking to establish the cognitive validity of the Aptis Listening Test.*

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

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527–535. Doi: 10.1037/0033-2909.112.3.527

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*, 319–342. Doi: 10.1111/j.1745-3984.2001.tb01130.x

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, *21*, 31–41. Doi: 10.1111/j.1745-3992.2002.tb00083.x

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73. Doi: 10.1111/jedm.12000

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

Knoch, U., Fairbairn, J., & Huisman, A. (2015). *An evaluation of the effectiveness of training Aptis raters online (VS/2015/001)*. London: British Council.

Lunz, M., Wright, B. & Linacre, J. (1990). Measuring the impact of judge severity on examination of scores. *Applied Measurement in Education, 3(4),* 331–345.

McCray, G., & Dunn, K. (2020). *Validity and usage of the Aptis Grammar and Vocabulary (Core) component.*

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., & Sheehan, S. (2010). *A Core Inventory of General English*. British Council / EAQUALS.

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., & Dunlea, J. (2015). *Aptis General Technical Manual Version 1.0*. London: British Council.

O'Sullivan, B., & 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. Doi: 10.1191/0265532202lt219oa

Shaw, S., & 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., & Schmitt, N. (2012). Lexical coverage and L1 and L2 listening comprehension: the same or different from reading comprehension? *Applied Linguistics, 34*(4), 1–24.

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

Weir, C. J., & 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.

Zheng, Y., & Berry, V. (2015). Aptis for Teens: Analysis of Pilot Test Data. *Technical report (TR/2015/004),* London: British Council.

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: Aptis task specifications: Aptis Grammar and Vocabulary component

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

Appendix C: Aptis task specifications: Aptis Listening component

1. MCQ A1
2. MCQ A2
3. MCQ B1
4. Multiple matching
5. Opinion matching
6. Double MCQ

Appendix D: Aptis task specifications: Aptis Reading component

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

Appendix E: Aptis task specifications: Aptis Speaking component

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

Appendix F: Aptis task specifications: Aptis Writing component

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

Appendix B: Aptis task specifications: Aptis Grammar and Vocabulary component

Task: Multiple choice sentence completion

Test	Aptis	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.					
Further task focus information	Each item will target a grammatical exponent from a specific level (A1–B2). A sentence (referred to as the stem) will be used to contextualise the targeted exponent. All elements of the stem and options will be constructed according to the categories specified in Features of the Input Text and Features of the Response (see below for details).					
Instructions to candidates	Presently no direct instructions. It is suggested that we add a generic rubric at the beginning of the Grammar part (not necessary to repeat for each item): *There are 25 items in this section. For each item, choose the best word or words to complete the sentence.*					
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
Lexical level: further criteria	All vocabulary used in the stem sentence must come from one level below the targeted grammatical exponent. For A1 and A2 grammatical targets, words in the stem come from K1, for B1 grammatical targets, vocabulary in the stem comes form K1–K2, etc. (See Guidelines on Adhering to Lexical Level).									
Grammatical level	The grammar of the stem sentence used to contextualise the targeted grammatical exponent should be from levels below that of the targeted exponent. For A1 and A2 grammatical targets, the grammar of the surrounding stem should be A1 exponents, for B1 targets, from A2 exponents, etc. (See guidelines on Adhering to Grammatical Level).									
Topic	Choose from topic list appropriate for the targeted level.									
Functions	Choose from the list of functional exponents for the targeted level.									
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 the list of grammatical exponents for the targeted level (e.g. for B2 tasks, choose grammatical exponents from the B2 exponent list). Note that some exponents are marked "not used as targets". These exponents should not be used as the targets for grammar items.					
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	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 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 Response						
Target	**Length**	1	**Lexical**	K1	**Part of speech**	Nouns, verbs, adjectives
Distractors	**Length**	1	**Lexical**	K1	**Part of speech**	Nouns, verbs, adjectives
Key information	Within sentence		Across sentences		Across paragraphs	
Extra criteria	1) All 5 targeted words and all of the bank of options must be the same part of speech. 2) All targeted synonym pairs will be generated from a finite list of synonym pairs. 3) The 5 distractors will be selected from the same K1 level and part of speech as the 5 targeted words.					
Presentation	Written		Aural		Illustrations/Graphs	

Task: Meaning in context

Test	Aptis	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 from a bank of 10 to complete each sentence. The correct word will be the most appropriate and plausible lexical choice for the context.					
Further task focus information	The sentence containing the gap should contain enough contextual information to secure the correct answer, and provide enough context for a competent speaker to predict the correct answer (or a range of plausible alternatives).					
Instructions to candidates	Complete each sentence using a word from the drop-down 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	
Lexical level: Further criteria	(See Guidelines on Adhering to Lexical Level for more information).										
Grammatical level A2	A1 Grammatical exponents (See Guidelines on Adhering to Grammatical Level).										
Grammatical level B1	A1–A2 Grammatical exponents (See Guidelines on Adhering to Grammatical Level).										
Topic	Topics from the list of topics for the targeted level.										
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 (see Genre list for more information).										

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	
Extra criteria	1) The target words should not be from the same semantic/lexical fields. 2) The distractors should be relevant to the targets. Each distractor should be relevant to 1 target. The relevance can be in terms of the semantic field/domain of activity of the contextualising sentence or the targeted word					
Presentation	Written		Aural		Illustrations/Graphs	

Task: Definition

Test	Aptis	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.					
Further task focus information	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. B1 is a transitional level, bridging the restricted field of activity open to Basic Users at A1/A2. From B1, learners become more independent, and an important part of that independence is utilizing the target language to acquire knowledge in the target language.					
Instructions to candidates	For each of the 5 definitions below, select the word that matches the definition from the drop-down 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
Lexical level: Further criteria	(See Guidelines on Adhering to Lexical Level for more information).									
Grammatical level	A1–A2 Grammatical exponents (See Guidelines on Adhering to Grammatical Level).									
Topic	Topics from the list of appropriate topics for B1.									
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			
Extra criteria	1) The target words should not be from the same semantic/lexical fields. 2) Each distractor should be designed to be relevant to 1 target, but capable of being ruled out by the definition.					
Presentation	Written	Aural	Illustrations/Graphs			

Task: Collocation

Test	Aptis	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.					
Further task focus information	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 that is required to correctly complete the task. A vocabulary item relevant to the level is being targeted to determine the depth of the test-taker's knowledge regarding that word. The collocation itself is not the target.					
Instructions to candidates	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 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 Response						
Target	**Length**	1	**Lexical**	K4–K5	**Part of speech**	Nouns, verbs, adjectives, adverbs
Determining collocation appropriacy	1) Consult the BYU–BNC resource for the targeted word. 2) Appropriate collocations should have a frequency of 10 or greater. 3) Appropriate collocations should have an MI of 3 or greater. 4) All other options in the bank should have a collocation frequency of 0 (zero) or 1 (one).					
Distractors	**Length**	1	**Lexical**	K1-K4	**Part of speech**	Nouns, verbs, adjectives, adverbs
Key information	Within sentence		Across sentences		Across paragraphs	
Extra criteria	1) The bank word selected to collocate with the target (i.e., be used immediately following the target) will be from a lexical level below the target (i.e. if the targeted word on the left is K5, the word to selected from the bank of options would be K4 or lower). 2) See criteria for determining collocation appropriacy above. 3) Set idiomatic phrases and sayings are not used. (.e.g. *apron + strings*, for which the most productive and likely usage is associated with the idiomatic expression "tied to your mother's apron strings"). 4) The approach to creating sets of targets and distractors is the same as *meaning from context* (i.e., targets should all come from different lexical fields, distractors should be related to a target in a one-on-one relationship). 5) Subjective, expert quality review will still be necessary to determine collocation appropriacy, and to avoid two possible answers. The collocation search noted above will not take into account cases when two words treated as not a possible pair by the frequency count, may actually collocate with intervening lexical items occurring in between the pair. This will need to be made explicit to item writers and quality reviewers to check.					
Presentation	Written		Aural		Illustrations/Graphs	

Appendix C: Aptis task specifications: Aptis Listening component

Task: MCQ A1

Test	Aptis	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.					
Further task focus information	The task focuses on identification of a specific word or number in a short message from familiar, everyday life situations, involving a speaker who is known to the intended listener. The task will NOT require the test-taker to imagine they are the intended listener.					
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	3-option multiple choice	**Items per task**	1			
Time given for part	Approx. 40 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 *(establishing propositional meaning/inferencing in Reading)*					
	Discourse construction *(building a mental model / creating a text level representation in Reading)*					

Features of the Input Text										
Length	30 seconds	**Words**	60–80	**Speed**	3.0 -3.5 syllables per second					
		Syllables	90 -105							
Accent	Standard British English speaker likely to be encountered in the UK.									
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	From topic list for A1.									
Text genre	Recorded telephone messages. The message may come from situations likely to occur in one of several domains (see above). In all cases, 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			
Extra criteria	1) The stem is the same question as in the instructions. 2) The targeted information will not be paraphrased. 2) The distractors are not be used in the input text. Only the targeted information will be heard in the text.					

Other features of the recording and task	
Other	A 3-second pause is inserted between the rubric and the input text.

Task: MCQ A2

Test	Aptis	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 must be a short question, e.g. *What is Arturo's phone number*?					
Presentation	Written	Aural	Illustrations/Graphs			
Response format	3-option multiple choice		**Items per task**	1		
Time given for part	Approx. 40 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 (establishing propositional meaning/inferencing in Reading)					
	Discourse construction (building a mental model / creating a text level representation in Reading)					

Features of the Input Text					
Length	30 seconds	**Words**	60–80	**speed**	3.0 -3.5 syllables per second
		Syllables	90-105		
Accent	Standard British English speaker likely to be encountered in the UK.				
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, 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 etc.).									

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				
Extra criteria	1) The targeted information will not be paraphrased. 2) The distractors will be used in the input text. 3) The targeted information may still be 1 word or a short phrase, but will involve understanding at the propositional level (e.g. how will they go to the concert?). The key information should require integrating simple, explicit information, with the links clearly signalled, across sentences (utterances).					

Other features of the recording and task	
Other	1) For dialogues, the speakers will always be 1 male and 1 female. 2) A 3-second pause is inserted after the instructions before the message begins.

Task: MCQ B1

Test	Aptis	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.					
Further 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 must be a short question (Example: *What is special about the painting?)*					
Response format	3-option multiple choice		**Items per task**	1		
Time given for part	Approx. 40 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 (establishing propositional meaning/inferencing in Reading)					
	Discourse construction (building a mental model / creating a text level representation in Reading)					

Features of the Input Text											
Length	30 seconds	**Words**	90–120	**Speed**	4.0 -5.0 syllables per second						
		Syllables	120-150								
Accent	Standard British English speaker likely to be encountered in the UK.										
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: Further criteria	The cumulative coverage should reach 95% at the K3 level. No more than 5% of words should be beyond the K3 level. (See Guidelines on Adhering to Lexical Level for more information).										
Grammatical level	A1–B1 Grammatical exponents (See Guidelines on Adhering to Grammatical Level)										
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 etc.).										

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			
Extra criteria	1) The targeted information will be paraphrased, and where appropriate/possible will be paraphrased. 2) The distractors will be used in the input text, and where appropriate/possible will be paraphrased. 3) The targeted information should require integrating information across utterances. The relationship between pieces of information will not be marked as explicitly as at A2, and the cohesion/links between information will utlilise referential links, substitution, ellipsis, to indicate the links between propositions.					

Other features of the recording and task	
Other	1) For dialogues, the speakers will always be 1 male and 1 female. 2) A 3-second pause is inserted after the instructions before the message begins.

Task: Multiple matching

Test	Aptis	Component	Listening	Task	Multiple Matching

Features of the Task						
Skill focus	Identifying factual information					
Task level (CEFR)	A1	A2	B1	B2	C1	C2
Task description	Identifying aspects of a topic and matching each aspect to a speaker. Listen to a short description to identify factual information.					
Instructions to candidates	The instructions will provide a context for the 4 speakers, the overall theme, and a task instruction. Four people are talking about X. Complete the table below. Example: *Four students are talking about their studies. Complete the table below.*					
Response format	Select correct answer from 6 options in drop-down list.			**Items per task**	4	
Time given for part	Approx. 40 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 (establishing propositional meaning/inferencing)					
	Discourse construction (building a mental model / creating a text level representation)					

Features of the Input Text					
Length	30 sec x 4	**Words**	70-90 x 4	**speed**	4.0 – 5.0 syllables per second (approx..)
		Syllables	115 - 125		
Accent	Standard British English speaker likely to be encountered in the UK.				
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: Further criteria	The cumulative coverage should be 95–100% at the K3 level. No more than 5% of words (i.e., 2 words) should be beyond K3. Main target information to be within K3 range.									
Topic	From topic list for B1. Overlapping content across inputs so there is at least one plausible distractor for each text.									
Text genre	Monologues are in the form of *vox pop* pieces. The genre is the same for each monologue in the task.									
Relationship of participants	4 monologues, each delivered by a different speaker. The speaker may or may not be known to the intended listener.									

Features of the Response						
Stem	**Length**	1-4 words	**Lexical**	K1-K2	**Grammar**	A1-A2
Presentation	Written	Aural	Illustrations/Graphs			
Options	**Length**	1–5 words	**Lexical**	K1–K2	**Grammar**	A1-A2
Presentation	Written	Aural	Illustrations/Graphs			
Key information	Within sentence	Across sentences	Across paragraphs			
Extra criteria	1) Input should contain B1 grammar components. 2) There should be information overlap across inputs. 3) Key target information should be paraphrased, even if isolated lexical items may occasionally be the same in the text and item. The candidate should not be able to answer the item purely through lexical matching. 4) There must be overlapping content across inputs so there is one plausible distractor for each text.					

Task: Opinion matching

Test	Aptis	Component	Listening	Task	Opinion Matching

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	The candidate listens to a dialogue between two speakers and identifies whose opinion matches the statement. The information targeted will require the integration of abstract ideas and propositions across an extended stretch of interaction.					
Further task focus information	The candidate can listen to the dialogue twice by pressing the play button.					
Instructions to candidates	Example: *Listen to two people discussing a social issue. Read the opinions below and decide whose opinion matches the statements below, the man, the woman, or both the man and the woman. You can listen to the discussion twice.*					
Response format	Identify who expresses each of the four given opinions: the male speaker, the female speaker, or both the male and female speaker.			**Items per task**	4	
Time given for part	Approx. 40 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 (establishing propositional meaning/inferencing)					
	Discourse construction (building a mental model / creating a text level representation)					

Features of the Input Text											
Length	120–140 seconds		**Words**	Approx. 400	**Speed**	4.5 – 5.5 syllables per second					
			Syllables	540 -630							
Accent	Standard British English speaker likely to be encountered in the UK.										
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: further criteria	The cumulative coverage should reach 95–100% at the K5 level. No more than 5% of words should be beyond K5. Main target information to be within K5 range.										
Topic	From topic list for B2.										
Text genre	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 professor and student, etc) The text should begin with a brief contextualisation. There will be some redundant information between sections of information targeted by the items.										
Relationship of participants	Dialogues: participants may be known to each other (friends, colleagues, teacher/student) or unknown (interviewer/interviewee, etc.).										

Features of the Response						
Presentation	Written	Aural	Illustrations/Graphs			
Options	**Length**	4–8 words	**Lexical**	K1–K4	**Grammar**	A1–B1 exponents
Presentation	Written	Aural	Illustrations/Graphs			
Relationship of participants	Dialogues: participants may be known to each other (friends, colleagues, teacher/student) or unknown (sales assistant/customer, public announcement etc.).					
Key information	Within sentence	Across sentences	Across paragraphs			
Extra criteria	1) Four opinions will be expressed in the dialogue in the order of the information in the text. 2) The opinions will be expressed by either the male speaker only, or the female speaker only, or by both the male and female speaker. 3) The opinions as they appear in the table should not appear verbatim in the text, but will be referenced through use of paraphrase and inference. 4) The targeted information should not be contained within a single sentence. The listener will be required to identify the targeted information across more than one sentence. 5) Opinions expressed should be plausible and balanced. Taboo topics to be avoided					

Task: Double MCQ

Test	Aptis	Component	Listening	Task	Double MCQ

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	The candidate listens to a monologue to identify two opinions/attitudes which are expressed by the speaker. The information targeted will require the integration of abstract ideas and propositions across an extended stretch of interaction.					
Further task focus information	This level targets more abstract information likely to be encountered in educational/public domains, and is designed to measure test-taker's ability to participate in these aspects of the TLU domain.					
Instructions to candidates	Example: *Listen to a woman talking on the radio about New Year's resolutions and answer the two questions below*					
Response format	3-option multiple choice asked as a question			**Items per task**	2	
Time given for part	Approx. 40 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 (establishing propositional meaning/inferencing)					
	Discourse construction (building a mental model / creating a text level representation)					

Features of the Input Text											
Length	80–100 seconds	**Words**	Approx. 300	**speed**	4.5 -5.5 syllables per second (approx)						
		Syllables	360-450								
Accent	Standard British English speaker likely to be encountered in the UK.										
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	K 2	K3	K4	K5	K6	K7	K8	K9	K10	
Lexical level: further criterai	The cumulative coverage should reach 95–100% at the K5 level. No more than 5% of words should be beyond K5. Main target information to be within K5 range.										
Topic	From topic list for B2.										
Text genre	Monologues: lectures, short speeches, short features on broadcast media, reviews on TV and radio, presentations in a work context.										
Relationship of participants	The speaker will be addressing an audience (either directly or remotely through broadcast media). The speaker may or may not have a relationship with the intended listener(s).										

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			
Extra criteria	1) The targeted information will be implied (not stated) by the speaker. 2) The targeted information should not be contained within a single sentence. The listener will be required to identify the targeted information across more than one sentence. 3) The distractors must be plausible and relevant to the content of the text, but may not be directly referenced in the text. 4) Distractors may be paraphrased as necessary and if appropriate. 5) The opinions will be phrased as complete stand-alone sentences. 6) The second item should target information at a holistic /discourse level using information in different parts of the text to ascertain the speaker's general opinion or attitude to the topic.					

Appendix D: Aptis task specifications: Aptis Reading component

Task: Multiple choice gap-fill

Test	Aptis	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.					
Further task focus information	The sentence containing the gap should contain enough contextual information to secure the correct answer, and provide enough context for a competent speaker to predict the correct answer (or a range of plausible alternatives). The task is presented as a text, but the level of comprehension targeted is A1, sentence level comprehension. Test-takers do not have to read beyond the stem sentence to fill the gap.					
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	35 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	40–50 words (including target words for gaps)					**Number of sentences (total)**	6			
Avg. sentence length	10–12 (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: further criteria	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	From topic list for A1. (For personal notes and letters etc. 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. The topic list is still relevant for identifying the range of possible points/information which might be mentioned).									
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			
Extra criteria	The distractors should not be able to be ruled out by the structural/collocation relationship to the word immediately before or after the target. The sentence containing the gap should contain enough contextual information to secure the correct answer, and provide enough context for a competent speaker to predict the correct answer (or a range of plausible alternatives).					
Presentation	Written	Aural	Illustrations/Graphs			

Task: Sentence re-ordering

Test	Aptis	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	Reorder two sets of jumbled sentences to form two short, cohesive texts. For each text, six sentences are presented, with the introductory sentence given first in the right order. The remaining sentences must be reordered to form a short text. The two texts appear on separate screens and are not related to each other.					
Further task focus information	The task tests A2-level comprehension, not higher-level discourse structure. The level of cohesion is restricted to linear, inter-sentential cohesion, so the order should proceed in a clearly linked order from the introductory sentence. Different types of cohesion should be exploited, including *reference (pronouns)*, *substitution and ellipsis*, *conjunction*, *lexical cohesion* (see glossary).					
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 (5) of jumbled sentences.					
Items per task	6 (The scoring algorithm recognises various permutations and awards a maximum of 6 marks for this task).					
Time given for part	35 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	80–90 words per text	**Text length**	6 (1 introductory sentence + 5 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: further criteria	All vocabulary should be from within the K1 and K2 levels (See Guidelines on Adhering to Lexical Level).									
Grammatical level	A2 Grammatical exponents (See Guidelines on Adhering to Grammatical Level).									
Readability	Flesch Kincaid of 4–6 (Cohmetrix advices using texts of 200 words or more for stable readability results. Because of the short nature of the A2 texts, this is an approximate guideline only.)									
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, etc.). 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?* and *Is the genre relevant to TLU tasks important for 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	
Distractors	All five sentences are required to form a cohesive text in combination with the introductory sentence. Each sentence is both target and distractor.						
Key information	Within sentence		Across sentences		Across paragraphs		
Extra criteria	Do not rely on only one type of cohesion to link all sentences. Try to use several types of cohesion across the text.						
Presentation	Written		Aural		Illustrations/Graphs		

Task: Opinion matching

Test	Aptis	Component	Reading	Task	Opinion Matching

Features of the Task						
Skill focus	Text level reading comprehension. Reading short paragraphs to comprehend the main ideas.					
Task level (CEFR)	A1	A2	B1	B2	C1	C2
Task description	Candidates read 4 short paragraphs giving information about 4 people's opinions on different aspects of a topic, e.g., travel, parental rules, school canteens, etc.. Candidates identify which of the four people are most likely to give certain statements.					
Further task focus information	Each paragraph is self-contained and centres on one individual. The paragraphs are linked through a common topic focus and contextual setting.					
Instructions to candidates	The instructions will provide a context in terms of a common topic focus for the 4 paragraphs, and a final task instruction. The people may (but not necessarily) have the same role, e.g., teacher, student.					
Response format	The same 4 named people from the drop-down list. *Four people were interviewed by a newspaper about a local park. Read the texts and complete the questions below.*					
Items per task	7 (each person to be the answer either 1 or 2 times)					
Time given for part	35 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	70-80 words per paragraph				**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: further criteria	The cumulative coverage should reach 95–100% at the K3 level. No more than 5% of words should be beyond K3. Main target information to be within K3 range.									
Readability	Flesch-Kincaid grade level of 6–8									
Topic	From topic list for B1.									
Text genre	Magazines, newspapers, Internet articles, online comments, (e.g. contextualised 'below the line' comments to an article. The texts are adapted to the level. Although not intended to be authentic, they should reflect features of relevant texts from the TLU domain. Relatively informal in tone but avoiding reliance on idioms. It should be possible to answer the questions: *Where would a reader be likely to see a text like this outside the test?* and *Is the genre relevant to TLU tasks important for General test-takers at B1 level?* Key information to be based on overall opinion and attitude.									
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						
Stem	**Length**	Maximum 10 words	**Lexical**	K1–K2	**Grammar**	A1–A2
Stem format	Each stem is phrased as a question: Who …? followed by a short statement. (e.g. *Who thinks the park is a nice place for families*?)					
Options					**Part of Speech**	NA
Key information	Within sentence		Across sentences		Across paragraphs	
Presentation	Written		Aural		Illustrations/Graphs	
Extra criteria	1) The target ideas or opinions will not be expressed directly in the text, but require textual inference, linking pieces of information expressed by one participant. This should not be across adjacent sentences where possible. 2) The ideas and information will be mainly concrete with limited abstract information and deal with familiar concepts and ideas relevant to B1 level learners in the TLU. 3) The wording of the item should avoid reliance on contrasting information (e.g.…more than…) or refer only to concrete information, avoiding lexical matching to answer the item successfully 4) Information overlap across paragraphs should be used. For example, an idea could be mentioned in two paragraphs, but the participants will have different opinions about it. 5) Of the four participants, three of them will express two targeted points of view (for a total of seven options).					

Task: Matching headings to text

Test	Aptis	Component	Reading	Task	Matching headings to text

Features of the Task

Skill focus	Expeditious global reading of 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.					
Further task focus information	The task is designed to elicit expeditious global reading of longer expository and argumentative texts relevant to the TLU domain for B2-level candidates of Aptis General. Test-takers are expected to be able to recognise the main idea and macro-propositions of each paragraph and integrate them into a discourse level representation.					
Instructions to candidates	Read the passage quickly. Choose the best heading for each numbered paragraph (1–7) from the drop-down 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	35 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; further criteria	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, etc.). It should be possible to answer the questions: *Where would a reader be likely to see a text like this outside the test?* and *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–K4	**Grammatical**	A1–B1
Distractors	**Length**	Up to 10 words	**Lexical**	K1–K4	**Grammatical**	A1-B1
Key information	Within sentence		Across sentences		Across paragraphs	
Extra criteria	1) All headings should avoid direct lexical overlap of key words in the paragraph they are intended to match. 2) Some ideas/concepts or key words in a target heading should overlap with ideas and information in more than one paragraph, but only represent the main idea/macro-proposition of one targeted paragraph (this is an ideal, but will be difficult to maintain across all seven target headings). Priority must be given to ensuring there is only one possible correct (best) combination for each heading/paragraph pair.					
Presentation	Written		Aural		Illustrations/Graphs	

Appendix E: Aptis task specifications: Aptis Speaking component

Speaking Task 1

Test	Aptis	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 3 spoken questions on personal topics. Each question is presented separately, and the candidate records his/her spoken response before the next question is presented.					
Task description: extra information	The task is designed to elicit short responses to spoken questions on familiar and concrete topics, and the rubric is phrased in the first person to approximate interaction with an interlocutor. Sets of 3 questions are generated by the system by randomly selecting 1 question each from 3 groups of questions designed to be comparable in difficulty.					
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.		
Timing of rating	Real time		After test event
Rater	Interlocutor	Rater present at test	Rater not present at test event

Speaking Task 2

Test	Aptis	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		B2	C1	C2
Task description	The candidate responds to 3 questions related to one picture prompt. The first question asks the candidate to describe a photograph. The candidate then responds to 2 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.					
Task description: extra information	The questions gradually increase in difficulty by expanding the focus from description of a concrete situation in a photograph to requiring the candidate to explain his/her opinions and elaborate on the topic. The rubric is phrased in the first person to approximate interaction with an interlocutor.					
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
Grammatical level	A1–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		
Topic	From topic list for A2/B1. The photograph will show 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.					
Timing of rating	Real time			After test event		
Rater	Interlocutor	Rater present at test		Rater not present at test event		

Speaking Task 3

Test	Aptis	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.					
Task description: extra information	The questions gradually increase in difficulty by expanding the focus from description of 2 photographs to comparison of aspects of the photographs, and finally providing his/her opinion and preferences with reasons and justifications. The rubric is phrased in the 1st person to approximate interaction with an interlocutor.					
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	
Grammatical level	A1–B1 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 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.		
Timing of rating	Real time		After test event
Rater	Interlocutor	Rater present at test	Rater not present at test event

Speaking Task 4

Test	Aptis	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).				
Task description: extra information	The task requires a long turn response in relation to abstract topics. The illustration is only for additional contextualisation of the topic and is not referred to directly in any of the questions.				
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
Grammatical level	A1–B1 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 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 6-point scale from 0–5. A B2-level performance is required to achieve score bands 3–4. A score of 5 is awarded for performances beyond B2 level.				
Timing of rating	Real time	After test event			
Rater	Interlocutor	Rater present at test	Rater not present at test event		

Appendix F: Aptis task specifications: Aptis Writing component

Writing Task 1

Test	Aptis	Component	Writing	Task	Task 1

Skill focus	Writing at the word or phrase level. Information to simple questions in a text message type genre.					
Task level (CEFR)	A1	A2		B2	C1	C2
Task description	The candidate answers simple questions. All responses are at the word or phrase-level. Each response will consist of responses to five questions.					
Task description: extra information	All tasks in the writing test are designed to build on common theme. Task 1 provides the initial setting in which the candidate provides basic information. The subsequent tasks will require increasingly longer pieces of writing eliciting different functions related to that theme. The task always has an example: *How are you? I'm fine, thanks.* The task then asks a selection of simple questions. Here are some examples: *What do you do? What did you do yesterday / last week? What's your favourite colour / sport? What's the weather like? How do you get to the shops / work?*					
Instructions to candidates	The instructions will clearly identify how to answer the questions.					
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. Three minutes is recommended for Task 1.					
Delivery	Pen and paper		Computer			
Response format	Word completion	Gap-filling	Form filling	Short answer	Continuous writing	
Intended genre	Text message type					
Writer / intended reader relationship	The reader will not be well known to the writer. The writing is transactional in nature and the reader is understood to be a new member of a club.					
Discourse mode	Descriptive	Narrative	Expository	Argumentative	Instructive	
Domain	Public	Occupational	Educational	Personal		
Nature of task	Knowledge telling		Knowledge transformation			
Functions targeted	Providing information (based on British Council EQUALS Core Inventory)					

Features of the Input / Prompt											
Description	Short questions with space for inputting short answer responses by the candidate.										
Number of categories	5										
Number of gaps	5										
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							
Information targeted	Information which is easily recoverable from memory and which an A1-level candidate is expected to be able to communicate.										

Features of the Expected Response				
Description	5 short gaps which can be filled by responses of 1–5 words.			
Length of response	Each gap can be filled by responses of 1–5 words.			
Lexis /grammar	K1 level lexis sufficient to complete task.			
Rating scale for task	A task-specific rating scale is used for the task. The rating scale is a 4-point scale from 0–3. 3 – Intelligible responses for all five questions. Test-taker achieves the task and answers all five questions. 2 – Three or four of the responses are intelligible. Errors impede understanding in one or two responses. 1 – One or two of the responses are intelligible. Errors impede understanding in three or four responses. 0 – No intelligible responses.			
Timing of rating	Real time		After test event	
Rater	Interlocutor	Rater present at test	Rater not present at test event	Automatic scoring
Weighting	Each task is weighted differentially to reflect the task demands and intended level. Task 1 contributes the least to the overall test score.			
Rating extra information	Each task for the same candidate is marked by a different rater. No one rater will mark more than 1 task for a single candidate.			

Writing Task 2

Test	Aptis	Component	Writing	Task	Task 2

Skill focus	Short written description of concrete, personal information at the sentence level.					
Task level (CEFR)	A1	A2		B2	C1	C2
Task description	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 as part of a form.					
Task description: extra information	The task builds on setting as Task 1, increasing the cognitive and linguistic demands by requiring sentence-level writing in response to a single 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
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		
Information targeted	The information targeted would be concrete, everyday and familiar information about the candidate, the candidate's personal experiences or surrounding, 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.					
Timing of rating	Real time	After test event				
Rater	Interlocutor	Rater present at test	Rater not present at test event	Automatic scoring		
Weighting	Each task is weighted differentially to reflect the task demands and intended level. Task 2 contributes fewer marks to the overall test score than tasks 3 & 4.					
Rating extra information	Each task for the same candidate is marked by a different rater. No one rater will mark more than 1 task for a single candidate.					

Writing Task 3

Test	Aptis	Component	Writing	Task	Task 3

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 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 background activity used in parts 1 and 2.					
Task description: extra information	The task builds on the same background setting as Tasks 1 & 2, but takes place within a social-media, interactive communication setting. The task increases the cognitive and linguistic demands by requiring a series of sentence-level responses to questions.					
Instructions to candidates	The instructions will clearly 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 other readers.					
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 (based on British-Council EQUALS Core Inventory. Note: describing hopes and plans is listed as B2 in the Core Inventory but when expressed in simple terms would be appropriate for a simple B1-level transfer of information.					
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 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 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 sentence level with sufficient cohesion.					
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 B1-level performance is required to achieve score bands 3–4. A score of 5 is awarded for performances beyond B1 level.					
Timing of rating	Real time	After test event				
Rater	Interlocutor	Rater present at test	Rater not present at test event	Automatic scoring		
Weighting	Each task is weighted differentially to reflect the task demands and intended level. Task 3 contributes fewer marks to the overall test score than task 4.					
Rating extra information	Each task for the same candidate is marked by a different rater. No one rater will mark more than 1 task for a single candidate.					

Writing Task 4

Test	Aptis	Component	Writing	Task	Task 4

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 in the prompt (management, customer services, etc.).					
Task description: extra information	The task builds on the same background setting as Tasks 1, 2 & 3. The task is designed to elicit responses demonstrating control of both informal and formal registers appropriate for different kinds of writing.					
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
Grammatical level	B1 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 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 6-point scale from 0–5. An B2-level performance is required to achieve score bands 3–4. A score of 5 is awarded for performances beyond B2 level.			
Timing of rating	Real time		After test event	
Rater	Interlocutor	Rater present at test	Rater not present at test event	Automatic scoring
Weighting	Each task is weighted differentially to reflect the task demands and intended level. Task 4 contributes the most marks to the overall test score.			
Rating extra information	Each task for the same candidate is marked by a different rater. No one rater will mark more than 1 task for a single candidate.			

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 1

Areas assessed: Task fulfilment and communicative competence

3 (above A1)	• Fully intelligible responses for all five questions. • Test taker completely achieves the task.
2 A1.2	• Three or four of the responses are intelligible. • Errors impede understanding in one or two responses.
1 A1.1	• One or two of the responses are intelligible. • Errors impede understanding in two or three responses.
0 A0	• No intelligible responses.

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.

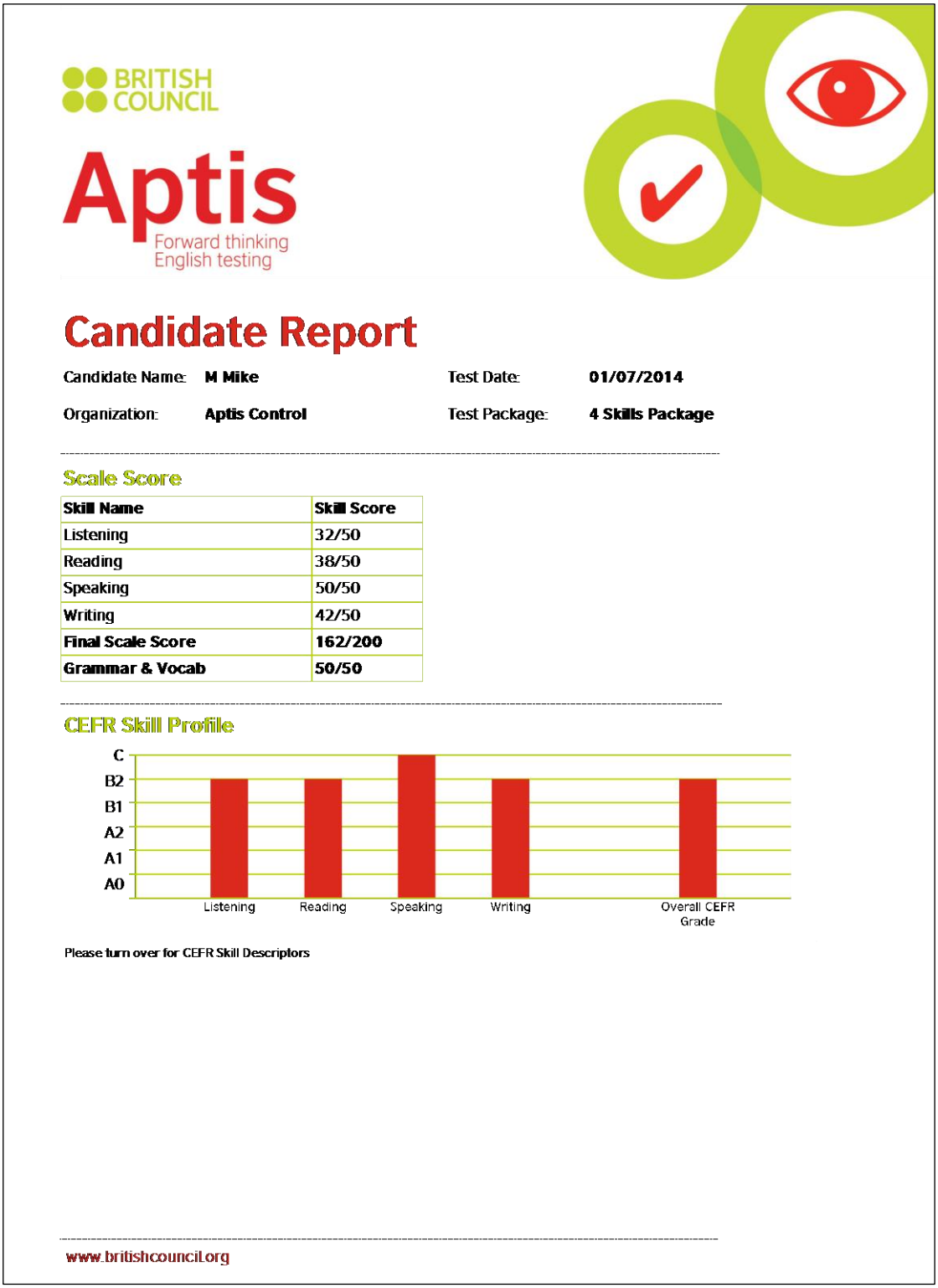
5 **B2 (or above)**	Likely to be above the B1 level.
4 **B1.2**	Responses to all **three** questions are on topic and show the following features • Control of simple grammatical structures. Errors occur when attempting complex structures. • Punctuation and spelling mostly accurate. Errors do not impede 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.

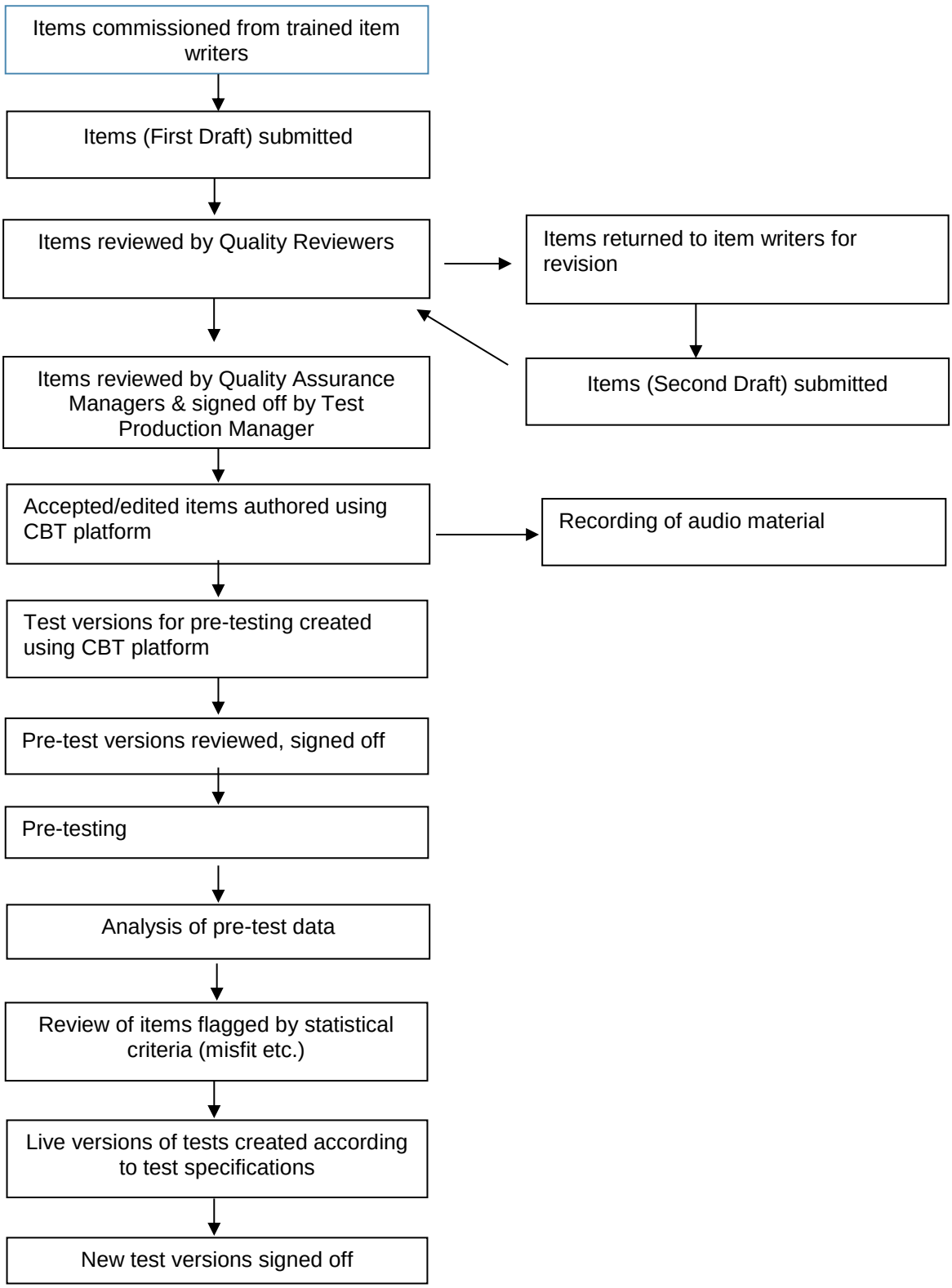
Speaking

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 an 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 2.2

Barry O'Sullivan, British Council
Jamie Dunlea, British Council
Richard Spiby, British Council
Carolyn Westbrook, British Council
Karen Dunn, British Council

ISSN 2057-7168