BRITISH
COUNCIL

ENGLISH LANGUAGE
ASSESSMENT RESEARCH GROUP

Technical Report

Aptis Scoring System Version 2.1

TR/2024/002

Karen Dunn
Assessment Research Group, British Council

ISSN 2057-7168

www.britishcouncil.org/aptis

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Acknowledgements

For their contribution to the content development and refining of early drafts, acknowledgements are made to Judith Fairbairn and David Akast of the British Council's Global Assessment team.

1. INTRODUCTION

A test's scoring system is one of three elements that lies at the heart of the test in the socio-cognitive model of validation (O'Sullivan, 2015). With this core role in mind, the Aptis scoring system has been developed to provide meaningfulness in the use and interpretation of scores for all stakeholders. This document explains the key features of the scoring approach for an audience with an interest in the mechanisms and technicalities of Aptis scores.

2. REPORT SUMMARY

This report describes the scoring and reporting processes in the Aptis suite of tests, as well as outlining the rigorous standards by which Aptis is constructed and aligned with the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001).

An overview of Aptis score reporting is given in Section 3. This is supported by information about setting and maintaining standards in Section 4 and the relationship between Aptis and the CEFR in Section 5. Information about the process by which Aptis is benchmarked to the CEFR and how cut scores are established is found in Section 6, with further details on CEFR level allocations for individual skill areas and overall L2 English ability in Sections 7 and 8 respectively. Finally in Section 9, there is a discussion on the use and interpretation of Aptis test scores, both on the numerical scale and as CEFR level.

3. OVERVIEW OF APTIS SCORE REPORTING

Aptis test performances for the four skills components – Listening, Reading, Speaking and Writing – are reported on two scales: as a numerical score and a CEFR level. Test performance in the Grammar and Vocabulary component, the "Core" component, is reported as a numerical score only. Candidates who sit all four skills components, plus Grammar and Vocabulary, are additionally awarded an overall numerical score and an overall CEFR level which summarises their L2 English language ability.

3.1 Scores and levels for individual skill areas

Scores are allocated for each skill based on candidates' achievement in the relevant test component. Writing and Speaking are marked by a team of trained examiners, while Listening, Reading and the Core component are marked using an automated system. Numerical scores for all components are reported on a scale of 0–50. The scale score is not equivalent to the raw score, or the number of correct items. As is common industry practice, raw scores are converted to scale scores for reporting.

For the skill components, a CEFR level is assigned according to the score obtained in each skill. Key points to note in this score allocation process are as follows:

1. Each skill component has designated set of score ranges which determine CEFR level allocation. The precise scores required to achieve a specified CEFR level differ for each skill, i.e. the range of scale scores that will achieve a B1 is different for Writing compared to Listening This means the numerical scores are not directly comparable between skill areas. See Appendix for CEFR cut scores for the main Aptis variants.

2. For a given skill component, if a candidate's score falls into the borderline area immediately below a CEFR cut-score boundary, CEFR allocation for that skill is refined based on their performance in the Core component. A strong Core performance will see them pushed up to the next CEFR level. It is possible therefore for two candidates to be awarded the same numerical score and a different CEFR allocation for the same component if they have scored differently in the Core component.

3.2 Overall CEFR level

Candidates completing a four-skills test are additionally allocated an *overall* numerical score, plus an *overall* CEFR level. The overall numerical score is the sum of the scores obtained in each skill. This is derived independently to the overall CEFR level which is calculated as a rounded average of the CEFR level achievement in each skill. The two scores are not directly comparable. This is described in more detail in Section 8.

4. APTIS AND STANDARDS

All Aptis tasks are written according to the published specifications (O'Sullivan and Dunlea, 2015) and are subject to a rigorous pre-testing procedure. When creating a new version, only new tasks which function in line with the existing standards will be included. The two key testing principles of reliability and validity are at the forefront of Aptis test design and the development of new versions. With respect to scoring validity, the theoretical fit of the way that performance is being assessed, the accuracy of scoring decisions made by raters, and alignment with the external performance standards outlined in the CEFR are all crucial elements. The latter is discussed in more detail in Section 5 below.

A test is reliable if it can be depended on to provide a consistent measurement of candidate ability levels, and a test is valid if it is measuring the ability it is designed to assess. The reliability of Aptis receptive skills components (Reading, Listening, and Core) are calculated at regular intervals using the Alpha statistic, and are invariably found to exceed the recommended threshold. Figures reported from the 2016–17 Annual Operating Report are shown in Table 1. Statistical assessment of test reliability gives an indication of what is known as the "internal consistency" of a test, in other words, whether all the items are working together to assess the same underlying ability. The most common means of reporting reliability is a Cronbach's Alpha statistic. The reported scale is between 0–1, with estimates closer to 1 representing a higher degree of reliability. A figure greater than 0.7 is generally considered to indicate that a set of test items are functioning as required.

Table 1: Cronbach's Alpha mean values for major Aptis variants 2016–17

Component	Aptis	Aptis for Teachers	Aptis for Teens	Aptis Advanced
Core	.92	.91	.90	.86
Listening	.84	.83	.82	.76
Reading	.91	.86	.85	.75

For the productive skills (Speaking and Writing) a measure of "inter-rater" reliability is given. This is based on the examiner performance on a series of "control items" within the system, and indicates how consistently the examiners are marking. Each examiner's workload will comprise approximately 5% control items, i.e. responses taken from a pool of candidate performances that have been scrutinised by senior examiners. All examiners are required to mark within tolerance of the agreed mark, and if not, they are required to undertake further training.

The control items are regularly refreshed so that an examiner will not encounter the same control item more than once. The inter-rater reliability is calculated based on the agreement of examiners on the marks awarded for the control items. For the 2016–17 Annual Operating Report, a mean inter-rater reliability value of .89 was recorded for Speaking and .85 for Writing for the Aptis general variant.

5. THE COMMON EUROPEAN FRAMEWORK OF REFERENCE FOR LANGUAGES (CEFR)

The CEFR is an ongoing project to provide a comprehensive framework for the description of language proficiency at clearly defined levels. The CEFR was first published in 2001 with the goal to provide "a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe" (Council of Europe, 2001). It was quickly adopted not only across Europe but in other regions of the globe as a transparent, accessible tool for educators and policy makers. One of the key objectives of the CEFR is to facilitate "transparency in testing and the comparability of certifications"[1]. As a part of this framework, there is a widely referenced set of language proficiency descriptors, setting out basis of language proficiency across six levels:

A1 – A2 (Basic User)

B1 – B2 (Independent User)

C1 – C2 (Proficient User)

[1] https://www.coe.int/en/web/common-european-framework-reference-languages/uses-and-objectives

The CEFR levels give guidance about language in use across the main language skills. These are fully elaborated in the CEFR (Council of Europe, 2001), with additional scales and descriptors in the Companion Volume (Council of Europe, 2018). Descriptive information, illustrative of the type of performance to expect at the different proficiency levels is given in a range of scales for communicative activities and strategies, including overall reading and listening comprehension, spoken and written production, and spoken and written interaction[2]. Scales for linguistic competencies address aspects of grammatical and vocabulary range and accuracy but only in very broad terms. These are deliberately under-specified in the CEFR, as the CEFR was intended as a common framework applicable to multiple languages. The CEFR makes clear that aspects of grammatical and lexical knowledge need to be fleshed out in local applications for different languages (Council of Europe, 2001, p. 33).

The range of CEFR levels it is possible to be awarded in Aptis depends on the test variant taken. Aptis General, Teachers and Teens do not differentiate at the higher end of the scale between C1 and C2 level because the tasks included in the test are designed and written to challenge English language learners with a range of abilities between A1–B2. If a candidate does exceptionally well in one of these Aptis test variants, it is clear that their proficiency level exceeds B2, and will be in the C range. It is recommended in these cases for candidates to sit the Aptis Advanced test, which includes targeted tasks that can discriminate between performances at C1 and C2 level.

See Table 2 for a summary of the CEFR level coverage across major Aptis variants.

Table 2: CEFR level coverage for major Aptis variants

	A0	A1	A2	B1	B2	C(1)[2]	C2
General	✓	✓	✓	✓	✓	✓	N/A
Teachers	✓	✓	✓	✓	✓	✓	N/A
Teens	✓	✓	✓	✓	✓	✓	N/A
Advanced	A0–A2[1]			✓	✓	✓	✓

[1] Aptis Advanced does not differentiate between A0–A2 level achievement.
[2] Only Aptis Advanced differentiates between C1 and C2 level achievements; other variants designate achievement above B2 as “C” level.

[2] The original CEFR documentation contains 57 different scales detailing communicative activities, strategies, and competencies, with further scales added in the Companion Volume.

6. ALIGNING APTIS WITH THE CEFR

Upon first development, standard setting exercises were carried out for Listening, Reading, Writing and Speaking skill components of Aptis to provide confident translations of numeric score achieved to a CEFR level that reflects the candidates' English language proficiency. This involved calling on a panel of expert judges, each bringing keen insight into indicators and measurement of L2 language proficiency. These judges were tasked with independently setting the cut scores representing the minimum score required for a candidate to achieve each CEFR level in each skill area. To achieve this, the team of judges followed four broad steps:

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

For further details of the full alignment procedure please refer to O'Sullivan (2015).

Aptis has benefited from being able to base test specification, task development, and rating scale development from the outset on the CEFR. It is important to note that standard setting itself, while essential, is only one part of a comprehensive alignment process. In addition, ongoing research continues to contribute to the validation of the CEFR alignment (e.g. Dunlea et al, 2018). The Aptis approach to best practice in project delivery across not only exams but all of the areas of activity of the British Council is driven by an evidence-centred approach.

6.1 CEFR levels and the Core component

As noted in Section 3 above, a CEFR level is allocated for each of the individual skill areas, but not the Core component. The Core component assesses candidates' grammar and vocabulary knowledge. Since this knowledge underpins all language skills, 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 underspecified elements of the framework. As described in the *Technical Manual* (O'Sullivan and Dunlea, 2015, p. 28), work on relating Core component to the CEFR is ongoing. The Core component does nonetheless play a role in the CEFR level allocation system for each skill component, as is elaborated in Section 7 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 candidates who do not complete this component.

6.2 Comparability with other tests

The CEFR provides a descriptive context that may be used to interpret the meaning and practical significance of scores on language tests. If a test developed by another organisation has undergone a similar rigorous process of alignment with the CEFR as Aptis and is testing the same language areas, then the CEFR level awarded for Aptis will be broadly comparable. However, as noted above, the CEFR includes numerous individual scales targeting different communicative activities, meaning that there is scope for a wide degree of variation in the opportunities a candidate may be given to display their skills across different tests.

When trying to interpret results from any test, and particularly if attempting to compare results from different tests, it is essential to consult the test specifications of each to understand clearly what aspects of language proficiency the test is designed to measure.

7. CEFR LEVEL ALLOCATIONS FOR INDIVIDUAL SKILL AREAS

In testing, one of the most important decisions is where to locate the cut scores, or thresholds between adjacent levels of achievement. Often the side of a cut score that a candidate falls can have immediate and significant impact on the life of a test candidate.

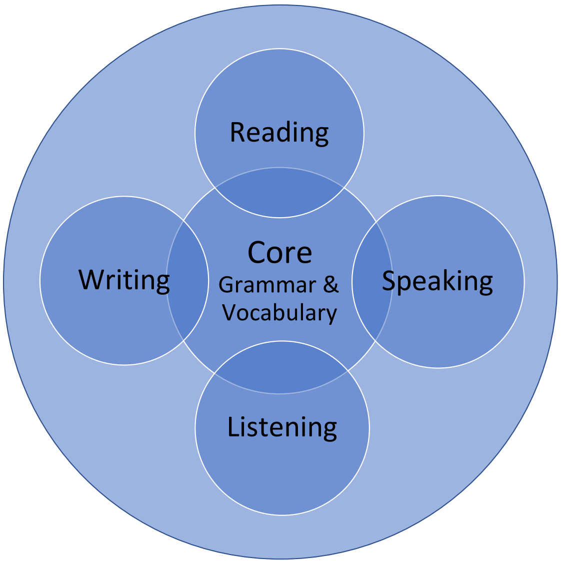
The decision as to where the cut scores will fall in an Aptis test component is reached via a process of standard setting (see Section 6 above), in which the decision is made regarding which score indicates that a particular level of English language proficiency has been reached for each successive CEFR level. Cut scores in Aptis are not the same across each of the skills components for each of the CEFR levels (see Appendix). This is because the content and design of the tests has been independently evaluated and aligned with the language descriptors set out in the CEFR. In order to accurately reflect a candidate's level of English language ability in any given skill area, each skill is addressed separately and the cut scores relevant for each CEFR level are based on the data and evidence collected in the alignment process. For Speaking and Writing skills, this reflects the quality of the language that is produced to achieve a score that meets each threshold. For Listening and Reading, this reflects the level of understanding indicative of a given CEFR level. This means that the exact point on the skill-specific score scale at which a person demonstrates a B2 level of proficiency, for example, will not be the same in the Reading component as it is for Listening.

In cases where a candidate's score for a given component falls in the borderline between two levels, Aptis uses score information from the Core component to refine the final CEFR level allocation. In this way, the Grammar and Vocabulary score feeds into the decision-making process for the awarding of CEFR levels across all skills: Listening, Reading, Speaking and Writing. This increases the fairness and accuracy of CEFR level allocation (McCray & Dunn, forthcoming).

The rationale for doing this is as follows:

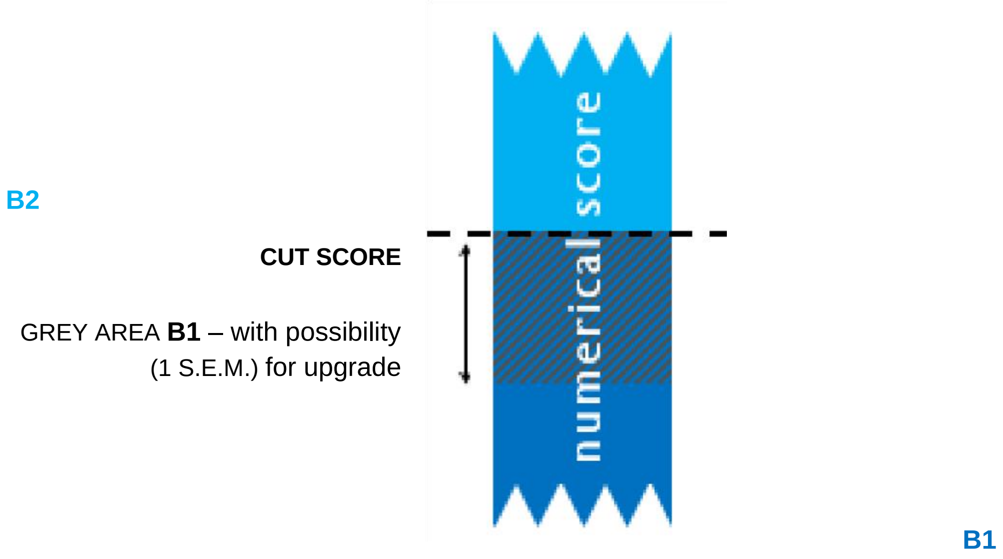
- In allocating a test score, there will always be a degree of error in the test's reflection of the candidate's true ability in any given skill. This is known as "standard error of measurement" in testing theory. The aim is to minimalise this error during test development, but it is impossible to eradicate fully.
- Grammar and vocabulary have been shown theoretically to be key processes in models of receptive and productive L2 language ability (e.g., Field, 2013; Grabe & Kaplan, 1996; Khalifa & Weir, 2009; Levelt, 1989), and empirically good predictors of language proficiency (e.g., Shiotsu, 2010). A solid foundation in grammar and vocabulary knowledge can therefore be viewed as relevant across all skill areas.
- Scores achieved in the grammar and vocabulary component provide a valuable additional source of evidence in finalising CEFR level allocation in the Aptis test for all skill areas, especially for the crucial thresholds between B1 and B2, B2 and C (McCray & Dunn, forthcoming).

Figure1: The Aptis Core component feeds into CEFR level allocation for all four skill areas



The process of determining CEFR level cut scores in Aptis is the same, regardless of the test variant taken (General, Teachers, Teens, Advanced). All candidates are required to sit the Core component, alongside the combination of skills tests they are entered for. The scale scores will be recorded and reported, however, if any of these scale scores fall just below a cut score threshold, i.e. within the margin of error, then the system refers to the candidate's Core component score to determine which side of the threshold is most likely to accurately reflect the candidate's ability. If the candidate has a strong performance in the Core component, they will be given the benefit of the doubt, and awarded the higher CEFR level. This process is illustrated in Figure 2.

Figure 2: Illustration of the "grey area" in which candidate CEFR level allocation is contingent on Core component performance



This process means that in the operational reality, it is fully possible for two candidates to achieve the same numerical score, e.g. 40/50 for the speaking test, yet for one of the candidates to be awarded a B1 and the other a B2. This is because 40/50 is an Aptis 'grey zone' score for Aptis General Speaking test. The candidate who was awarded a B2 will have performed better in the Core component than the candidate awarded a B1.

8. CEFR LEVEL ALLOCATIONS REPRESENTING OVERALL L2 ENGLISH ABILITY

The “overall” CEFR level provides a summary of the candidate’s L2 English language ability. This level is only awarded if a candidate completes all four skills components. In order to calculate this overall level, information about the performance in each of the four skill areas is essential. If overall CEFR levels were to be calculated from different skills combinations, i.e. from a 3-skills or 2-skills package, then comparability would not be maintained across different testing occasions and between candidates. The overall CEFR allocation is calculated as a rounded average of the four CEFR levels achieved in each of the skill areas. This is shown in Table 3.

Table 3: Illustration of overall CEFR level calculation

Listening	Reading	Writing	Speaking	Overall
A2	A2	A2	B1	**A2**
A2	A2	B1	B1	**B1**
A2	B1	B1	B1	**B1**

8.1 Relationship between the overall CEFR and the overall numerical score

Since the overall CEFR level allocations and the overall numerical scores are independently calculated, there is not a *direct* relationship between the two. While the overall numerical score is a straightforward total of the scores for each of the four skill areas (out of a maximum possible 200), the overall CEFR calculation is based on the four CEFR levels allocated. These individual CEFR levels are based on any adjustments necessary to reflect a strong performance in the Core component, as described in Section 7; however, the overall CEFR level calculation does not take into account whether a candidate achieved at the lower or higher end of the score range. This means that it is possible for two candidates to receive a wide range of overall numerical scores and the same overall CEFR allocation. Indeed, it is also technically possible for a candidate to receive a lower overall numerical score than another candidate but in fact achieve a higher overall CEFR level. See Table 4 for a comparison of some overall CEFR level and overall numeric score combinations.

Table 4: Illustrative combinations of overall numerical score and overall CEFR level allocations

Listening		Reading		Speaking		Writing		Overall	
36	B2	38	B2	34	B1	44	B2	**152**	**B2**
46	C	32	B1	43	B2	46	C	**167**	**B2**
42	C	44	C	41	B2	40	B2	**167**	**C**

9. INTERPRETING CEFR LEVELS AND NUMERICAL SCORES AWARDED TO CANDIDATES

Any given CEFR level for an individual L2 skill covers a broad range of abilities and competences. While test scores can demonstrate a candidate's level, it is important to remember that several performance descriptors exist within one CEFR level and it takes time to master all of them. This means that when measuring candidate progress on a course, it is unrealistic to expect a candidate to increase their language ability even by one CEFR level within a limited period of time.

This section of the report provides guidance on making valid score comparisons.

9.1 Individual skill areas

CEFR levels awarded for individual skills can be used to glean a profile of candidate ability. For example, it might be useful to know whether a candidate's receptive (listening and reading) skills are more pronounced than their productive (writing and speaking) skills. This can be achieved by comparing CEFR level achievement in each of the skills. It is not however valid to compare the numerical score across skill areas. For example, a candidate achieving 36 in both Listening and the Writing components cannot be said to have identical levels of ability in these two skills, this is because of the different alignment of each skills component with the CEFR (see Section 7).

In comparing ability *within a single individual skill area*, it is however a valid approach to refer first to the CEFR level and then to the numerical score for a more detailed comparison. So, for example, in comparing candidates who both achieved B2 in the Speaking component, it is valid to refer to the numerical score each candidate achieved on this component to establish which is the strongest at this skill. The numerical scores for individual skills can also be used to give an indication of improvements in Aptis test performances at the component level by the same person across different sittings in cases where an increase of a full CEFR level might not be possible. Equally, this information can be used to investigate the benefit of different teaching interventions for groups of students, since it is more nuanced than the CEFR level information.

9.2 Overall English language performance

When comparing the *overall* English language ability of candidates, it is most informative to refer to the overall CEFR level awarded upon completion of a four-skills test, rather than the total numeric score out of 200. Because of the different calibration of each of the skills components against the CEFR, candidates with the same total score could be awarded a different overall CEFR level (see Section 3.1). The overall CEFR level is intended however to give a broad overview of a candidate's English language ability.

REFERENCES

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

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

Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing: an applied linguistic perspective*. London: Longman.

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

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA, US: The MIT Press.

McCray, G., & Dunn, K. (Forthcoming). *Validity and Usage of the Aptis Grammar and Vocabulary (Core) Component*.

O'Sullivan, B. (2015). *Aptis Test Development Approach*. Retrieved from https://www.britishcouncil.org/sites/default/files/tech_001_barry_osullivan_aptis_test_-_v5_0.pdf

Shiotsu, T. (2010). *Components of L2 Reading: Linguistic and Processing Factors in the Reading Test Performances of Japanese EFL Learners*. Cambridge: Cambridge University Press and Cambridge ESOL.

Appendix: CEFR cut scores for main Aptis variants

The cut points used in scoring for the main Aptis test variants on the Common European Framework of Reference are given below. The figures represent the starting point of each level on the 0–50 scale for each component.

Aptis General (revised 2020)

	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

Aptis Advanced

	B1	B2	C1	C2
Listening	14	24	38	50
Reading	16	28	40	50
Writing	22	37	44	50
Speaking	22	37	44	50

Aptis for Teachers (revised 2023)

	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

Aptis for Teens

	A1	A2	B1	B2	C
Listening	6	12	26	40	45
Reading	6	12	24	38	46
Writing	6	18	26	40	48
Speaking	4	16	26	41	48

BRITISH COUNCIL
APTIS TECHNICAL REPORTS

ISSN 2057-7168

9 772057 716005 >

The British Council is the United Kingdom's
International organisation for cultural relations
and educational opportunities