BRITISH COUNCIL

ENGLISH LANGUAGE
ASSESSMENT RESEARCH GROUP

CHINA'S STANDARDS OF ENGLISH LANGUAGE ABILITY (CSE): LINKING UK EXAMS TO THE CSE

TECHNICAL REPORT ON LINKING UK EXAMS TO THE CSE

VS/2019/003

Jamie Dunlea, Assessment Research Group, British Council
Richard Spiby, Assessment Research Group, British Council
Sha Wu, National Education Examinations Authority (NEEA)
Jie Zhang, Shanghai University of Finance and Economics
Mengmeng Cheng, National Education Examinations Authority (NEEA)

BRITISH COUNCIL VALIDATION SERIES
SERIES EDITOR: RICHARD SPIBY AND CAROLYN WESTBROOK
ISSN 2398-7979

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. STEERING GROUP AND WORKING GROUP MEMBERS

Steering Group

Barry O'SULLIVAN	British Council
Nick SAVILLE	Cambridge Assessment English
Jianda LIU	Guangdong University of Foreign Studies
Lianzhen HE	Zhejiang University
Wenxia ZHANG	Tsinghua University

Working Group

Jamie DUNLEA	British Council
Richard SPIBY	British Council
Sheryl COOKE	British Council
Jane LLOYD	Cambridge Assessment English
Jing XU	Cambridge Assessment English
Gad LIM	Cambridge Assessment English
Maggie DUNLOP	Cambridge Assessment English
Sha WU	National Education Examinations Authority (NEEA)
Jie ZHANG	Shanghai University of Finance and Economics
Shangchao MIN	Zhejiang University
Mingwei PAN	Guangdong University of Foreign Studies
Hongwen CAI	Guangdong University of Foreign Studies
Weiqiang WANG	Guangdong University of Foreign Studies
Linlin CAO	Guangdong University of Foreign Studies
Manman GAO	Anhui University
Mengmeng CHENG	National Education Examinations Authority (NEEA)
Hao ZHANG	Tsinghua University
Fan YANG	National Education Examinations Authority (NEEA)

2. EXECUTIVE SUMMARY

2.1. Overview

This technical report presents the findings of a comprehensive research project to develop a methodology for appropriately linking UK examinations to China's Standards of English Language Ability. This Executive Summary presents a brief overview of the background and methodology and the main linking results achieved for the two tests included in this project, Aptis and IELTS. For more comprehensive information, a detailed description of the CSE itself is presented in Section 4, the methodology employed is described in Section 5, and the key results and recommendations are described in Section 7 of the body of the report.

2.2 Background

On 6 December 2016, Mr Chen Baosheng, China's Minister of Education, and Ms Justine Greening, the UK Secretary of State for Education, signed *Action Plan under the UK–China Partners in Education Framework* at the 4th Meeting of the China–UK High-Level People-to-People Dialogue in Shanghai, China. Strand 2 of the document states that to promote communication and alignment between the English proficiency standards of each country, China and the UK will conduct collaborative research on linking various UK English language tests to China's Standards of English Language Ability. The National Education Examinations Authority (NEEA), Ministry of Education, China and the British Council (BC) were appointed respectively to implement this joint program.

The collaborative research is a joint endeavour to strengthen UK–China cooperation, and support a golden era in UK–China relations. The IELTS and Aptis tests were chosen as the pilot projects to build a demonstration of best practice in terms of investigating the relationship between the CSE and language examinations.

2.3 Methodology

2.3.1 Joint research team

A professional joint research team of more than 20 people was set up to implement the research program. The team members include representatives and researchers from NEEA, Chinese academic institutions, the British Council, and Cambridge Assessment. Their experiences and expertise cover a wide range of areas related to language assessment, i.e. standard development, test development, standard setting, teacher training, statistics, as well as project management.

A clear team structure was established (see below) and the responsibilities were well defined for each role, to ensure the implementation effective and efficient. The Steering Group oversees the academic design and implementation of this goal. The Working Group is responsible for carrying out the necessary research activities. Project coordinators and event managers were appointed among NEEA and British Council China to manage the internal and external communication and the logistics of the research activities.

2.3.2 Research framework and agenda

The linking research is a process of validation. To make a robust claim, multiple sources of evidence are needed. With the aim to demonstrate the best practice, the research team employed integrated approaches to make the research solidly grounded.

The research went through three main stages: 1 – construct evaluation; 2 – standard setting (split into three phases); and 3 – external validation. The process includes building a construct definition of the tests, rough descriptor analysis and test content analysis, collecting expert panel judgements, as well as gathering actual test data from students, in addition to teacher judgements as external validation.

A detailed research agenda was drawn up to provide guidelines for the two-year research program. A series of events were planned and clear outcomes identified.

2.4 Summary of results

The Working Group has completed the Construct Phase and the Expert Panel Standard Setting Stage and in addition, it has undertaken an External Validation study as part of Phase 3. The detailed findings of the study are presented as a series of recommendations in Section 7.4 in the main report. The key alignment cutscore recommendations are presented below.

IELTS and Aptis related recommendations

The outcomes of the standard-setting panels, in conjunction with other sources of evidence collected as part of the linking project, have been shown to offer an accurate and consistent estimate of the cutscores relating China's Standards of English Language Ability (CSE) to the IELTS and Aptis tests.

Recommendation 1

The cutscores suggested by the Working Group based on the standard-setting panel results should be adopted with immediate effect. These cutscores may be updated based on the rollout of the CSE and of further planned research into the link between IELTS/Aptis and the CSE.

IELTS	CSE 4	CSE 5	CSE 6	CSE 7	CSE 8
Listening	5	6	6.5	7.5	8.5
Reading	4.5	5.5	6	7	7.5
Speaking	5	5.5	6	6.5	7
Writing	4	5	6	7	7.5
Overall*	**4.5**	**5.5**	**6**	**7**	**8**

* IELTS reports a profile and an overall band score which is derived from averaging the band scores on the profile. This table reflects this approach.

Aptis	CSE 3	CSE 4	CSE 5	CSE 6	CSE 7
Listening	14	21	29	37	43
Reading	16	26	35	42	46
Speaking	21	29	37	43	47
Writing	22	31	39	45	50

* Aptis reports a profile and an overall score. The overall CEFR/CSE level is estimated by first calculating the CEFR/CSE level independently for each of the four skills and then averaging the CEFR/CSE levels. This table reflects this approach.

3. INTRODUCTION

3.1 Background

This project was carried out as part of an MOU signed in December 2016 between the National Educational Examination Authority (NEEA) and the British Council to work collaboratively on research related to China's Standards of English Language Ability. As a central focus of that research, clause 2.1 of the agreement proposed the following objective: develop a research agenda to investigate the feasibility of linking various UK English language examinations to the CSE.

Building on that initial agreement, a Steering Group was formed in May 2017 to oversee the academic design and implementation of this goal. The Steering Group consisted of representatives from NEEA, Chinese academic institutions, the British Council Assessment Research Group, and Cambridge Assessment. A Working Group was also established with researchers from the same organisations to carry out the necessary research activities. This technical report is intended to provide a comprehensive overview of the project, and to provide clear documentary evidence detailing the process of linking and the results obtained.

3.2 Scope and goals

The Steering Group proposed to focus on two English as Foreign Language (EFL) examinations, Aptis and IELTS (the tests are described further in Section 6). The results of the linking study are important in and of themselves: they are of interest to educators and researchers interested in gaining insights into the descriptions of proficiency contained in the CSE levels and in understanding which parts of the CSE levels can be considered relevant to the two examinations. However, the project was intended from the outset to do more than link two specific tests, aiming also to develop a coherent, theory-based methodological framework for investigating the relationship between any language examinations and the CSE. As such, it is intended to be a demonstration of best practice in terms of investigating the relationship between the CSE and language examinations. While standard setting to determine cutscores for particular CSE levels is central to the project methodology, the framework described here is a multi-method approach which posits that multiple strands of evidence need to be collected in order to fully understand the relationship between an examination and a set of standards. Crucially, this evidence includes construct definition in order to clearly make the logical argument for what aspects of proficiency are targeted by the tests and how these relate to the CSE levels.

The project was carried out in stages (the methodology is described in more detail in Section 5). The Steering Group and Working Group first evaluated the construct definition of the tests to determine whether linking the tests would be feasible and meaningful. Following this qualitative evaluation, the Working Group carried out an initial pilot standard-setting panel with the listening components of the two tests, using test-centred standard-setting methods. This pilot was intended to trial the proposed methodology, after which the procedures were reviewed in light of the results of the initial study. Separate panels were then planned for the reading, speaking, and writing components. In addition, a separate strand of data collection was carried out by administering the tests to a large sample of students and having those students judged in relation to the CSE by their teachers. This stage of data collection used the Contrasting Groups, examinee-centred standard-setting method. Additionally, members of the Working Group also carried out a more detailed construct definition exercise separately, evaluating the content of the tests task-by-task using an evaluation template based on the socio-cognitive model, and mapping each task to particular descriptors and scales within the CSE.

The aim of this report is to provide a detailed description of these various strands of data collection, and the results. The intention is that researchers will be able to replicate these data collection procedures. Additionally, we suggest that best practice in linking exams to the CSE should contain all of these various strands of data collection, and should not rely on any one aspect, including standard setting itself, to justify a claim of a link to the CSE.

4. CHINA'S STANDARDS OF ENGLISH LANGUAGE ABILITY

4.1 Context

With an increasing awareness of the need to scale English learners' proficiency, language educators, teaching practitioners and policymakers in China agreed that a unified proficiency scale was urgently needed to describe learners' performance and streamline their competence across different educational stages and different regions in the Chinese EFL context. In 2014, the State Council of China issued a document entitled *Deepening the Reforms on the Educational Exams and the Enrolment Systems*. One pressing task, as highlighted in the document, was to develop a foreign language assessment framework to improve the quality of language tests, enhance the communication between teaching, learning and assessing, and thus raise the overall effectiveness and efficiency of foreign language education in China. For this purpose, the National Education Examinations Authority (NEEA), endorsed by the Ministry of Education, China, initiated a nationwide project to develop an English language proficiency scale, known as the China's Standards of English Language Ability (CSE), which set out to: (1) define and describe the English proficiency of English learners in China; (2) provide references and guidelines for English learning, teaching and assessment; and (3) enrich the existing body of language proficiency scales for alignments on a global basis (Liu, 2015).

4.2 Theoretical groundings

Bachman (1990) and Bachman and Palmer (1996) proposed the Communicative Language Ability (CLA) model, where language ability is characterised as "consisting of both knowledge, or competence, and the capacity for implementing, or executing that competence in appropriate, contextualised communicative language use" (Bachman, 1990, p. 84). The CLA model not only includes organisational competence in its traditional sense, but also incorporates strategic competence, which is regarded as serving not just a compensatory function, and thus, to a certain extent, alludes to Canale's (1983) refined model. More importantly, the CLA model recognises the roles of cognitive strategies and pragmatic competence, together with their impact on the realisation of communicative competence. As a whole, the model is theoretically sound, empirically validated, and is considered the state-of-the-art representation of language ability (Alderson & Banerjee, 2002).

Contingent upon their purpose, language proficiency scales may have different orientations. The Common European Framework of Reference (CEFR), which was published by the Council of Europe in 2001, provides reference descriptors for six main proficiency levels (and three "plus" levels). The CEFR adopts an action-oriented approach to the description of language use. This approach highlights whether and, if so, how language users integrate various language skills in performing particular activities in social, public, educational, and workplace settings. The action-oriented approach views users and learners of a language primarily as 'social agents' who have tasks (not exclusively language-related) to accomplish in certain circumstances, environments and fields of action. Therefore, language use is treated as being composed of the actions performed by individuals and social agents as they develop both general and communicative language competences (Council of Europe, 2001). This orientation proved to be appropriate in European contexts, given the role that English plays in those circumstances. Europeans, many of whom use English as a second language, have many more opportunities to communicate in the language, either spoken or written. However, in comparison with their European counterparts, English learners and users in China are more prone to use English in an educational context. Thus, in order to justify the real use of the target language (Bachman & Palmer, 2010); one principle of CSE development is to prioritise how English is used in China's EFL context (Yang, 2015).

The CSE takes a use-oriented approach to the description of language ability based on the CLA model and the educational needs of Chinese English learners. In the CSE, language ability is defined as the ability to comprehend and express information that learners exhibit when they apply their

language knowledge and world knowledge, and the strategies to perform language use tasks in a variety of contexts (Liu & Han, 2018). The CSE treats language ability as a type of dynamic cognitive activity, instead of an abstract and static system of rules. A salient feature of the CSE is the Chinese EFL context. The use-oriented approach focuses on the description of language use, covering the typical language use behaviours of different levels of Chinese English learners and users. The language ability of Chinese English learners and users is reflected by their participation in various activities involving language use, whether those activities are interactional or non-interactional in nature. Such a conception regards language ability as consisting of language comprehension, language expression, and mediation. The CSE has both an overall description of the language ability of Chinese English learners and users and specific descriptions matching their different levels.

4.3 Components

As is illustrated in Figure 1, the core of the CSE reflects an overarching notion of language ability, with which language knowledge and strategies co-function in performing a language activity. This mechanism sits consistently within the CLA model, where communicative success depends upon the language knowledge learners and users resort to, as well as the strategies they employ.

Language ability can be further divided into language comprehension (listening and reading), language expression (oral and written) and mediation (translation and interpretation). Of these, the description of mediation ability is ground-breaking in that, to date, there have been no proficiency scales that have dealt with this ability comprehensively[1]. In congruence with the main functions that communication serves, different sub-abilities deal with a plethora of texts, including narrative, descriptive, expository, argumentative, instructive and interactional texts. The two-headed arrow between translation/interpretation and functions/texts means English learners or users need to operate through two channels: the source and the target languages, both of which are manifested by texts of various communicative functions. Therefore, in order to streamline the framework across different sub-ability scales, the CSE developers laid down the “four-layer framework”, meaning that the description of language ability is structured in a hierarchical system. Language ability stands in the top layer, beneath which are language comprehension, language expression and mediation. The third layer comprises listening, reading, speaking, writing, translation, and interpretation. All these six sub-abilities are described based on six functions or text types, which make up the fourth layer. Global scales for overall language ability and each sub-ability are provided, as well as the sub-scales for all the functions specific to each sub-ability mentioned above.

To specify what constitutes language knowledge, the CSE developers mainly referred to the CLA model (Bachman & Palmer, 2010). Language knowledge consists of organisational knowledge and pragmatic knowledge. The former can be further broken down into grammatical knowledge and textual knowledge; the latter includes functional knowledge and sociolinguistic knowledge.

Apart from language knowledge, strategies can be divided into planning, execution and appraising/compensation. Different language sub-abilities are heavily involved in all of these. It is worth noting that the names of strategies related to different language sub-abilities can differ. For example, as derived from the umbrella term of appraising/compensation, the specific name for the writing sub-scales may be editing/proofreading, whilst that for the speaking sub-scales may be repairing.

[1] A recent publication of additional scales for the CEFR has a strong focus on what is also referred to as mediation, but the interpretation differs from the focus taken by the CSE. For more information, see: the *CEFR Companion Volume with New Descriptors* published by the Council of Europe: https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Figure 1: Components of the CSE (China's Standards of English Language Ability)

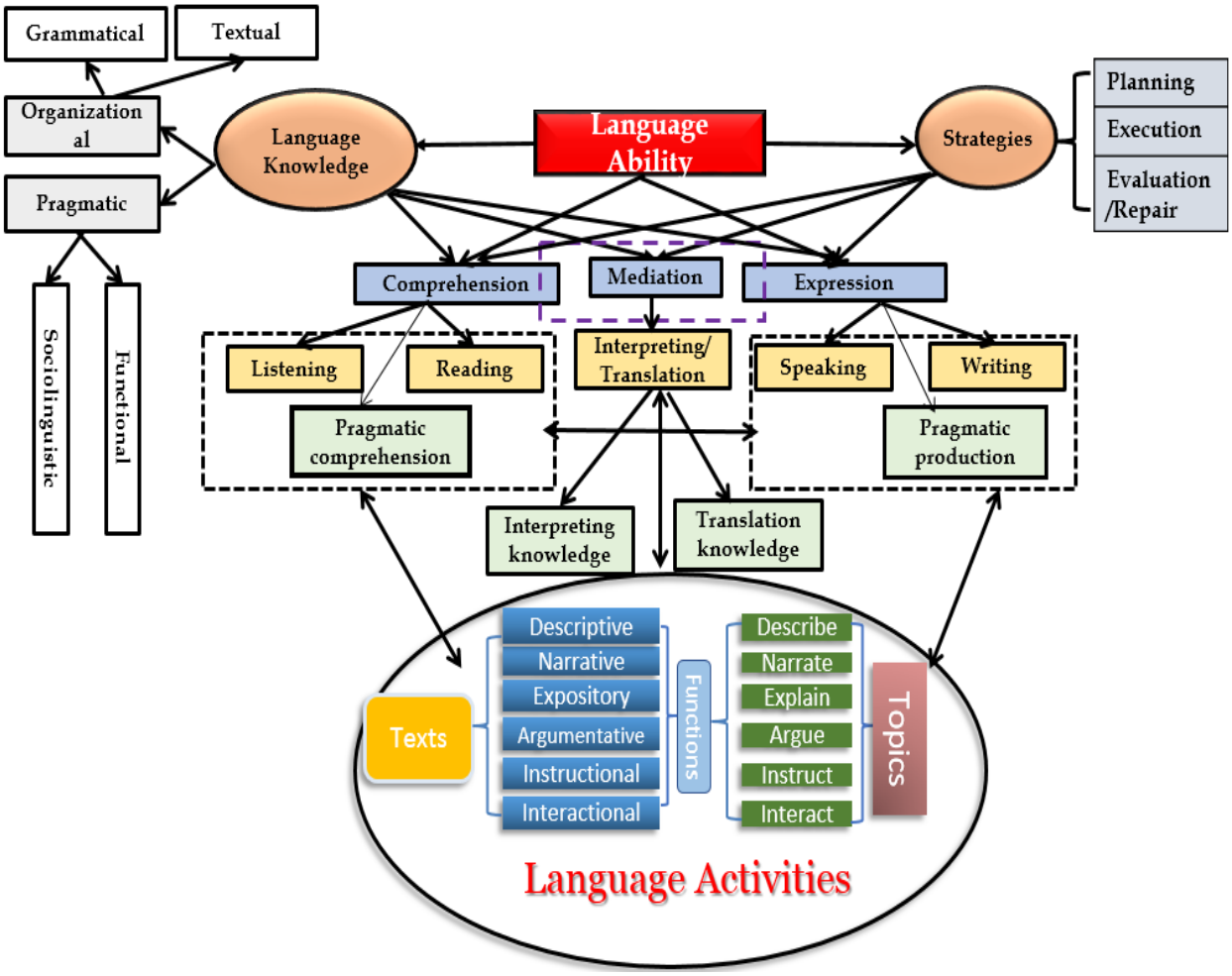
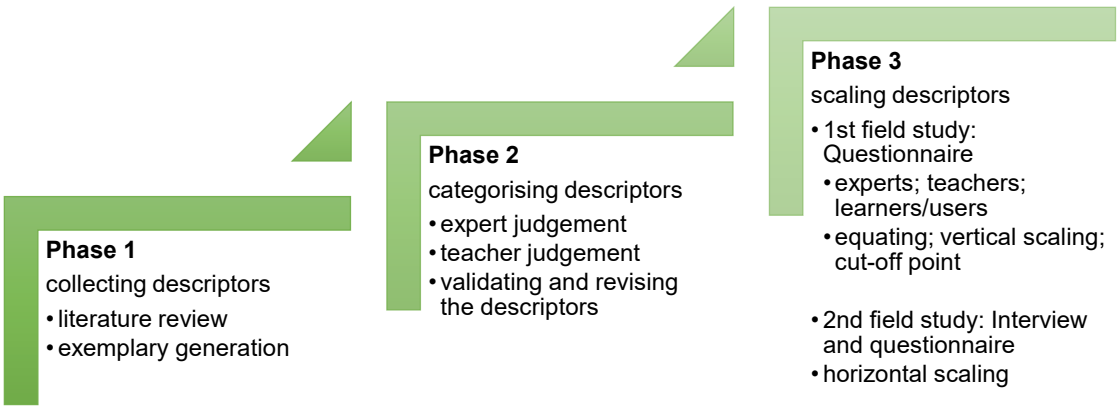


Figure 2: The CSE development procedure



4.4 Development and validation

The project of developing the CSE was launched in June 2014 and completed at the end of 2017. Figure 2 outlines the CSE development procedure. As illustrated in Figure 2, the development of the CSE can be briefly divided into three phases. The first phase primarily dealt with collecting descriptors, which derived from not only a wealth of literature but also a database of descriptors generated by students and teachers of different educational levels. In the second phase, the CSE developers, based on expert and teacher judgements, conducted trial validation on a working-group basis. During this process, the developers removed duplicate descriptors, blended similar descriptors and categorised descriptors into the framework of the CSE, as explained above. The last phase was composed of two field studies for the finalisation of scaling. In the first field study, all the refined descriptors were randomly spread into different sets of questionnaires, which were then administered to language education experts and frontline teachers, as well as learners/users. They reported the extent to which their students (if the participants were teachers) or they themselves (if the participants were learners/uses) could perform in relation to each descriptor provided. Based on the results, statistical analyses were conducted to determine the cut-off points of each proficiency level. The second field study, which was smaller in scale, attempted to elicit responses from teachers of various educational stages to the same descriptors, so that horizontal scaling could be done for the calibration of the cut-off points.

Based on the composite analysis of the research results mentioned above, CSE descriptors were scaled into 9 levels: CSE 1, CSE 2 to CSE 9, and were arranged in an ascending order from lower proficiency levels to higher ones. These levels correspond to elementary (CSE 1–3), intermediate (CSE 4–6) and advanced (CSE 7–9) English language proficiency. However, it is worth noting that the CSE sets out to describe the progression of general English language proficiency regardless of learners' educational background. The alignment is intended to promote understanding and communication among the public, language educators and teaching practitioners.

Three guidelines provided a consistent thread through the whole process of descriptor screening and revision. First, **each descriptor should take the form of a "can-do statement"**. In other words, what is described should point to learners' or users' accomplishments rather than their weaknesses. Caution was also taken in using hedging and degree adverbs, such as 'comparatively' and 'in general' for scaling purposes. That meant the descriptors could stand alone. Descriptors that were long or structurally complex were also revised so that CSE users would not be at a loss as to the focus an individual descriptor. Ambiguity, vagueness, atypical language activities and linguistic jargon were all avoided wherever possible.

Second, **the intended construct of an individual descriptor should be unique**. This was particularly true for the descriptors in the translation and interpretation sub-scales. If more than one ability were included in a descriptor, this could give rise to misunderstandings among CSE users.

Third, **each descriptor should follow a three-element model** (Pearson Standards and Quality Office, 2014) as shown in Table 1 and outlined below:

- **performance:** the language operation itself (e.g. can answer the telephone)
- **criteria:** the intrinsic quality of the performance, typically in terms of the range of language used (e.g. using a limited range of basic vocabulary)
- **conditions:** any extrinsic constraints or conditions defining the performance (e.g. with support, if spoken slowly and clearly)

Table 1: Three-element model of descriptors

Descriptor	Performance	Criteria	Condition
Can extensively and coherently elaborate on his/her views on academic or professional topics. (Speaking Oral Argumentation Level 7)	Can elaborate on his/her views	extensively and coherently	on academic or professional topics
Can briefly describe changes to familiar people or surroundings. (Writing Written Description Level 4)	Can describe changes	briefly	familiar people or surroundings

As such, the element of performance, which stipulates the 'doing' with the English language, is required in all descriptors. In comparison, the other two elements are optional, given their role of adding or removing constraints.

5. METHODOLOGY

5.1. Linking examinations to a framework

5.1.1 Overview

There is an extensive body of literature within the field of educational measurement on investigating the relationship between an examination and a set of standards, including establishing cut-off points on examinations which can be considered representative of particular levels of proficiency. The CSE, however, is not merely a set of standards but is expected to function more like a reference framework. In that regard, the closest analogue would be the Common European Framework of Reference for Languages (CEFR). While the CSE is designed for a particular context of use, and there are important differences between the CSE and CEFR, there are also important similarities. Since its publication in 2001, a large body of work has been established in linking examinations to the CEFR. As such, reference was made to this body of work to draw on the experience of examination providers linking their tests to the set of proficiency standards in the CEFR. The intention was not to prioritise one set of standards or to suggest that the CSE needs to make reference to the CEFR. As noted, they are clearly distinct. However, drawing on the research and experience accumulated over the last two decades has allowed us to expedite the process, avoid pitfalls already experienced in other studies, and distil best practice into the comprehensive framework for linking we posited as a goal of this project form the outset.

An important document which has informed much of the work linking exams to the CEFR is the *Manual for Linking Examinations to the CEFR* (Council of Europe, 2009). All major studies linking exams to the CEFR have made reference to the Manual or to the earlier pilot version of it (Council of Europe, 2003). What is noteworthy about the Manual is its view of linking, not as a one-off activity, but as a process of collecting convergent validity evidence. In particular, it lists five steps in the process of building an argument to justify a claim of linkage to the CEFR: familiarisation, specification, standardisation, standard setting, and validation.

Two of the steps, **familiarisation and standardisation**, are about ensuring that users are aware of, and have a shared understanding of, the framework. These are requirements that should go without saying, but being stated highlights that people are involved and central to the linking process, and therefore they should possess certain qualities and knowledge.

Three of the steps are explicitly about linking the test to the CEFR.

Specification refers to the qualities of the test being linked to the framework. This stage emphasises that sufficient validity evidence for the exam itself is a prerequisite for any linking. In addition, frameworks such as the CSE are extensive, covering multiple levels across multiple skill domains, and illustrated by thousands of descriptors. No examination would be capable of targeting more than one part of the framework. The specification stage helps make clear which part and which levels of the framework are relevant to the test. Thus, any linking would only be valid as it relates to those parts and those levels.

Standard setting is described as being "at the core of the linking process" (Kaftandjieva, 2004, p. 1). It is defined as "the proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more states or degrees of performance" (Cizek, 1993, p. 100). In the case of linking to the CEFR and to the CSE, such a number is the point on a test score scale at which a test-taker can be considered to have demonstrated the level of proficiency described in one of the CEFR or CSE levels. Because different exams test different skills in different ways, different methods for standard setting are necessary. A number of these are discussed in Kaftandjieva (2010).

Validation refers to "the body of evidence put forward to convince the test users that the whole process and its outcomes are trustworthy" (Council of Europe, 2009, p. 90). This can be internal or external to the test and can involve a wide range of methodologies. The important thing to note is the focus on process and on having a collection of convergent validity evidence.

Convergent validity evidence is emphasised here because the literature on standard setting accepts different outcomes on different standard-setting exercises as being inevitable. Standard setting is not seen as an exercise in determining a "true" objectively-existing cutscore, but as a values-driven enterprise with a desired policy-related end (Camilli, Cizek & Lugg, 2001; Cizek, 1993; Kane, 1998; Zieky, 2001), and so different cutscores are to be expected. Indeed, Cizek and Bunch (2007) recommend against multiple methods because of the difficulty of resolving the different scores that will be generated. Nonetheless, Kane (2001) recommends multiple methods as a powerful form of validity evidence. The issue becomes how to establish the degree of difference that can be tolerated and the degree of similarity which would be required to interpret the results as convergent evidence. The mechanism for doing this can be provided by the construct definition phase, which should establish some broad *a priori* claims about the level of proficiency the test is targeting in relation to the framework in question. Convergent evidence – that will be different but interpretable as reasonable difference supporting the *a priori* claims – thus becomes a realistic expectation from the use of multiple methods, and Kane's recommendations become a powerful source of validity evidence. Dunlea (2015) demonstrates how such a mechanism can be developed in practice in a study linking a large-scale set of proficiency tests in Japan to the CEFR.

In the case of standards frameworks, however, it is not policy preferences being dealt with, but levels of proficiency and what people can do with a language. To wit, the CEFR and CSE both include thousands of descriptors of what people can do with language. If the standard says that someone "can understand with ease virtually everything heard", then a score of zero on a listening test cannot be a legitimate cutscore relating to that standard. There are thus cutscores that are simply incorrect, and cutscores that are closer to or matching the level under consideration.

In addition, multiple exams link themselves to reference frameworks such as the CSE. If cutscores that are not comparable are deemed acceptable, it will result in a situation where it is easier to obtain a level on one exam compared to another exam. Not only would this be unfair, but one of those results would clearly be spurious and not valid.

It should be clear from the above that standard setting to reference frameworks is a qualitatively different activity from usual, requiring a different approach to the treatment of standard-setting outcomes (Lim, Geranpayeh, Khalifa & Buckendahl, 2013). Convergent outcomes from multiple sources of data should be expected as evidence of valid linking outcomes.

5.1.2 Standard setting and the CEFR

The most extensive evidence-based approach to establishing links between an exam and such a framework resides largely within the field of work addressed by *standard setting* (see below for definitions and an overview of standard-setting research and methodology). Standard setting is, in essence, an attempt to provide a body of evidence, both quantitative and qualitative, to support the judgements by experts in mediating the interaction between the descriptions of expected / required performance in a particular area of activity at different levels of ability, and the actual operationalisation of those descriptions in test items which yield various forms of feedback, such as test scores. In much of the original literature for standard setting, which was largely developed and refined in the United States, the purpose has been to establish defensible decisions for setting pass marks on examinations, for example for graduation from a program of study, or for certification in a particular area of professional activity. The quality and degree of ability required are usually presented as verbal descriptions with varying degrees of specificity. These descriptions, when framed with the intention of informing cut-off decisions on examinations, are often referred to in the literature as *performance level descriptors* (PLDs). As noted above, this area of standard setting has strong policy-driven imperatives, and decision-makers setting such cutoffs often need to take into consideration a great deal of information outside the test and the performance level descriptors themselves.

Since the publication of the CEFR in 2001, however, an extensive range of studies has been produced which have adapted and applied standard-setting methodology for the purposes of linking examinations specifically to this descriptive framework of proficiency. In this case, the CEFR was not developed with any one examination or certification program in mind, and was from the outset intended to be agnostic in terms of specific teaching and assessment methodologies. The development of the CSE itself has taken the valuable experience of the CEFR into account, and indeed CEFR descriptors have been built into the extensive quantitative data collected in order to scale the CSE descriptors. As such, the application of standard setting in relation to the CEFR provides the most important, targeted, and relevant body of literature which will be useful for informing procedures for this pilot study to link examinations to the CSE.

As already noted, an important document which has informed much of the linking in relation to the CEFR has been the *Manual for Linking Examinations to the CEFR* (2009), produced by the Council of Europe. Within this document, it clearly states that carrying out linking procedures based on standard-setting procedures itself is of no meaning without first establishing the construct relevance of the examination to the content of the illustrative scales of proficiency which sit at the heart of the CEFR. It further emphasises that sufficient validity evidence for the exam itself is a prerequisite for linking. Standard setting does not itself provide validation for an exam or sufficient evidence of a link between that exam and the framework in question. It is an important, indeed necessary, part of establishing that link, but not sufficient. The recommendations section will address this in more detail, with specific recommendations for how to achieve this in practice. The following literature review focuses on providing a background to standard-setting methodology, particularly in relation to the CEFR, as this remains a centrally important part of establishing evidence for a link, and is an integral part of the procedures employed as a part of this study.

5.2. Overview of literature on linking examinations to standards

5.2.1 Overview

The following literature review, as mentioned above, draws heavily on the use of standard setting in relation to the CEFR. The CEFR provides an invaluable case study and a bank of published studies that describe the development of a framework of proficiency made explicit through verbal descriptions following a clear set of protocols to achieve consistency in those descriptions across skill areas and across levels. At the same time, attempts were made to go beyond expert judgement and to incorporate actual empirical information in the scaling of those descriptors to proficiency levels. The framework itself did not presuppose any particular test as an exemplification of its performance descriptions. Linking exams, which in the early stages of its development was done mostly retrospectively, as the exams pre-dated the CEFR, required the development of a clear set of procedures which would allow for the collection and evaluation of evidence that could support or refute a test developer's claims of a link. Researchers turned to standard setting as a crucial step in this process. The CSE has built on the wealth of knowledge and experience and published research that has been accumulated in the almost 17 years since the CEFR's publication.

The CSE has been developed for a particular context, China, and has drawn on new and different approaches and innovative methodologies in its development. In its sheer scale, it clearly dwarfs the initial development of several hundred descriptors calibrated and published as the illustrative scales as a part of the CEFR. Nonetheless, the overlapping features make reference to the experiences of researchers linking exams to the CEFR an invaluable resource. These features include a comprehensive approach to developing and scaling descriptions of proficiency into a coherent framework, and then aligning examinations to the framework that operationalise the performance level descriptions contained within it. The following section draws heavily on a review of the literature in Dunlea (2015), to contextualise the recommendations made in Section 5 against a background of best practice in linking examinations to a descriptive proficiency framework.

5.2.2 Linking examinations to the CEFR

The CEFR was published by the Council of Europe in 2001 following a 10-year development process (Morrow, 2004; North, Martyniuk & Panthier, 2010). As North (2007) notes in reference to the full name of the CEFR: "Assessment is in third place; the language testing profession is a service industry to support teaching and learning". Nonetheless, Morrow (2004, p. 8) recognises the importance of the scales to the framework's descriptive system, leading him to refer to the Common Reference Levels as being "at the heart" of the CEFR. These levels and the Illustrative Scales which define them were developed in two major projects carried out in Switzerland in 1994 and 1995 (North, 2000; North & Schneider, 1998). According to North (2000), what distinguishes the development of the CEFR from previous descriptive scales of proficiency, such as the American Council for the Teaching of Foreign Languages (ACTFL) scale, is the use of Rasch analysis to empirically validate the allocation of descriptors to difficulty levels. The calibrated descriptors are used to define six broad levels of language proficiency across a total of 54 separate scales describing communicative activities, strategies, and communicative language competences. Figueras et al. (2005, p. 266) describe the development of a manual in response "to the need for guidance to assist examination providers in relating their examinations to the CEFR". This resulted in a Preliminary Pilot version of a manual being published by the Council of Europe in 2003, and a subsequent revised edition in 2009. All major studies linking exams to the CEFR have made some reference to one or both versions of the Manual. The Manual (Council of Europe, 2009) lists five steps in the process of building an argument to justify a claim of linkage to the CEFR: familiarisation, specification, standardisation, standard setting and validation. The Manual is supported by a series of Reference Supplements dealing with technical issues related to linking examinations to the CEFR, including Reference Supplement B: Standard Setting by Kaftandjieva (2004).

Kaftandjieva (2004, p. 1) describes standard setting as being "at the core of the linking process". Standard setting is described by Cizek (1993, p. 100) as: "the proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more states or degrees of performance". In the case of linking to the CEFR, such a number is the point on a test score scale at which a test-taker can be considered to have demonstrated a level of proficiency described in one of the CEFR levels.

5.2.3 Standard-setting methods

Among the many standard-setting methods available to practitioners, the Basket Method has been widely employed across Europe in relation to linking to the CEFR (Kaftandjieva, 2009, 2010). However, in the wider international context, particularly in the United States, the Angoff method, or modifications collectively referred to as the Modified Angoff Method, is often cited as the most frequently used (Cizek & Bunch, 2007; Cohen, Kane & Crooks, 1999). It is also one of the most widely researched, with papers comparing it to other methods (Bowers & Shindoll, 1989; Livingston & Zieky, 1989) or investigating modifications of the Angoff method (Clauser et al., 2009; Hurtz & Auerbach, 2003; Norcini et al., 1987). Although the Angoff method has been criticised as placing too great a cognitive burden on participants (Cizek & Bunch, 2007), studies have shown it to be robust (Plake, Impara & Irwin, 2000) and less prone to statistical bias than other methods (Reckase, 2006). Zieky (2001) has also refuted claims that the judgement task is too cognitively demanding for standard-setting panellists. The Modified Angoff Method has not only remained one of the most widely used methods, but has been employed in a number of studies linking exams to the CEFR (e.g. Tannenbaum & Wiley, 2005, 2008; O'Sullivan, 2008; O'Sullivan, 2015b).

Standard setting has a long history of use in educational measurement in the United States (Kaftandjieva, 2010; Papageorgiou, 2010). However, in relation to the situation in Bulgaria, Kaftandjieva (2010, p. 23) describes the most commonly used methods for setting cutscores and passing standards on exams as: "tradition, authority and the Goldilocks method", with the latter referring to an arbitrary process of setting a cutscore such as 80% simply because "70% is too little and 90% is too much". However, it would not be unreasonable to suggest that a similar lack of familiarity with standard-setting methodology was also common in other European countries prior to the introduction of the CEFR. It was indeed the perceived lack of familiarity with procedures for setting cutscores and linking examinations which led to the production of the Manual (2003, 2009) and Reference Supplements (2004). A similar process of growing familiarisation with the principles of standard setting in conjunction with exposure to the CEFR can also be noted for Japan (see Section 5 for details of the familiarity of participants in standard-setting panels in this study).

Kaftjandjieva (2010, p. 29) gives a comprehensive account of standard-setting methods, describing 62 documented methods, but cautioning that even this list "is not complete". Cizek and Bunch (2007), Hambleton et al. (2000), Kane (1998), and Livingston and Zieky (1982) also provide useful overviews and descriptions of the most prominent methods. These methods are often separated into one of two categories – student centred or test centred – based on a classification system originally suggested by Jaeger in 1989 (Kane, 1998; Kaftandjieva, 2010).

Kane (1998, p. 131) describes the two approaches in the following way:

> In the test-centered methods, the judges review the tasks or items in the test and decide on the level of performance on these tasks that would indicate attainment of the performance standard…In the examinee-centered methods, performances of real examinees are evaluated relative to the performance standard, and the test scores of these examinees are used to set the cutscore. For example, in the borderline-group method, the judges identify examinees who just meet the performance standard and the cutscore is set equal to the median score for these examinees.

As noted, these two broad categories lend themselves to different skill areas, with examinee-centred methods often being used for productive skills tests, such as writing. However, the two broad groups of methods also offer an important source of collecting the kinds of convergent evidence from multiple methods that Kane (2001) recommends. Indeed, Kane specifically recommends employing both test-centred and examinee-centred standard-setting methods as a powerful source of validation. The methods utilised here for linking with the CSE have followed that recommendation in that there were two important phases of standard setting. One stage of standard setting employed an expert panel approach, with test-centred methods used for setting standards on the receptive skills of Listening and Reading, while an examinee-centred method was used for the panels focusing on Speaking and Writing. In a follow-up stage, a separate examinee-centred approach, known as the Contrasting Groups Method, entailed the collection of actual test score data from a large number of examinees. These examinees were also classified by trained teachers, who made reference to the CSE in doing so. The actual methods employed in each of these stages, with some adaptations specifically designed for this project, are described in more detail below.

5.2.4 Selecting judges for panels

It is important to note that all forms of standard setting involve some form of human judgement (Cizek & Bunch, 2007; Kane, 2001), and so the selection of judges is a crucial part of the process. In terms of the criteria for selecting judges, or raters, to participate in standard-setting panels, Jaeger (1991, p. 4) suggested that “expert judges should be well experienced in the domains of expertise we demand of them”. In terms of the number of judges, Raymond and Reid (2001) note a wide range in the literature, ranging from “admissions” of 5 to recommendations of 15 to 20. Hurts and Hertz (1999, p. 885) applied generalisability theory to eight studies using the Angoff Method and concluded: “10 to 15 raters is an optimal target range.”

5.2.5 Criteria for evaluating standard setting

A number of criteria has been suggested for evaluating the results of standard setting (Cizek et al., 2004; Cizek & Bunch, 2007; Hambleton, 2001; Kaftandjieva, 2010). Cizek and Bunch (2007, pp. 59–63) describe three main categories of evidence. Procedural validity evidence involves a description of the processes employed, including the training for participants; the degree of correspondence of the procedures to the requirements of the methods used, and also includes feedback from participants. Internal validity evidence looks at the accuracy and consistency of the results of the standard-setting methods used, including the degree to which participants converge toward a common standard over the course of standard setting rounds. External validity evidence includes comparison of results obtained from other standard-setting methods and other sources of information.

In terms of strengthening the plausibility of results obtained from standard setting, Kane (2001, p. 75) recommends replicating standard setting with different methods, suggesting that using different methods and participants “would provide an especially demanding empirical check on the appropriateness of the cutscore”. Cizek and Bunch (2007) take the opposite view, warning that there is no consensus methodology for reconciling the different cutscores likely to be generated by different methods.

This study has taken the position that the use of multiple methods can, in fact, be an important approach to external validation. Various approaches can be taken to ameliorate the concerns of Cizek and Bunch (2007), as described in Dunlea (2015). It is now widely accepted that different standard-setting methods will derive different cutscores (Cizek, 2001; Zieky, 2001). Indeed, Kaftandjieva (2010) also notes that different cutscores will be obtained if standard setting is replicated using the same method. Cizek and Bunch (2007, p. 63) also describe the reasonableness of the decisions made as an important criterion for evaluating the decisions made through standard setting. We take the perspective that evaluating the reasonableness of the decisions requires multiple data collection procedures to triangulate and evaluate from multiple perspectives. The problem Cizek and Bunch (2007) note when using multiple standard-setting methods, in terms of how to decide which cutscore is actually the “right” cutscore, is actually misleading. Despite our attempts to provide more robust

quantitative precision to our measurement instruments through IRT, etc., in the end, we are attempting to operationalise constructs which themselves remain works-in-progress in terms of language proficiency and acquisition. These constructs are then mediated through verbal descriptions in proficiency frameworks which, regardless of the strength of calibration of the descriptors, remain to a certain extent open to interpretation. We suggest then that the collection of multiple perspectives does not muddy the waters, but instead is the only way to establish confidence in the final claims of linking performance on the score scale of an examination to levels on a proficiency framework.

As noted above, the multiple stages necessary for a comprehensive, defensible linking methodology actually provide the mechanism for resolving the inherent differences in cutscores that multiple methods bring. The first stage, construct definition, allows some *a priori* assumptions to be made regarding the proficiency levels targeted by an exam. If there is relevance between the exam and the CSE (or other framework in question, such as the CEFR), the construct definition phase will allow a logical argument to be made which would identify and describe not only which proficiency levels are relevant to the exam, but also which parts of the framework the exam is relevant to. As noted above, no exam would, or possibly could, target all of the aspects and levels in the CEFR or the much larger – in terms of number of descriptors and categories – CSE. Once such *a priori* claims are made, it is possible to create a mechanism for identifying what a reasonable level of difference in cutscores would be that would still imply the cutscores provide supporting evidence for the *a priori* claims of relevance derived from the construct validation phase. In effect, explicit *a priori* criteria should be set which the researchers would consider to be evidence of convergent results from multiple methods and which would support the *a priori* claims. These explicit criteria then provide evidence for either justifying and defending those *a priori* claims or refuting them. These criteria thus provide a mechanism for identifying reasonable differences in standard-setting results from multiple methods, and thus provide a way of operationalising Kane's (2001) call to use both test-centred and examinee-centred methods as a powerful form of validity evidence. Dunlea (2015) provides clear exemplification of how this process can work in practice.

While adopting a clearly documented, principled approach to collecting and analysing data can inform cutscore decisions, Cizek and Bunch (2007) caution that the results "are seldom, if ever, purely statistical, psychometric, impartial, apolitical, or ideologically neutral activities". However, Cizek and Bunch (2007) also emphasise that decisions taken within the context of educational measurement always involve, to some degree, evaluative judgements by those tasked with making those decisions. Standard setting does not remove that burden or the difficulties inherent in carrying out those responsibilities. There will be no magic statistical procedure, technique or software application which will remove the need for principled decisions to be taken in relation to setting cutscores. Indeed, the validity and validation of a test itself is often framed as not an absolute decision, but a matter of degree established through a thorough evaluative argument (Messick, 1989). Standard setting should certainly be viewed in this light, and decisions should be made within a clear framework of reference and an understanding of the goals and contextual constraints under which the process is carried out.

5.3 Overview of recommendations for a comprehensive methodology

5.3.1 Principles for a framework on linking to the CSE

Given the discussion above, some principles can be derived that serve as a basis for linking to the CSE:

- linking to the CSE is a process composed of multiple steps
- linking to the CSE must involve proper consideration of the construct
- linking to the CSE must involve appropriate methods resulting in defensible, convergent outcomes.

In addition, the people involved in the linking process need to be suitably familiar with, and inducted into, the workings of the CSE.

5.3.2 Recommended design and procedures for linking exams to the CSE

Consistent with the principles outlined in the previous section, it was determined that the methodology for linking tests to the CSE should consist of several stages.

Stage 1 – Construct evaluation: In order to investigate the construct targeted by the exams, a model of language test validation is essential. It was determined that the socio-cognitive model of language test validation (Weir, 2005) should be employed, as: (a) it is in consonance with the model of language ability reflected in the CSE; (b) it has been used extensively in the development and description of a large number of exams; and (c) it has been proven to facilitate the interpretation of evidence in a range of studies.

Stage 2 – Standard Setting: In order to identify scores related to relevant CSE levels, the Modified Angoff Method and Analytical Judgement Method were selected. The Angoff method in particular is the most widely used internationally, including in studies linking exams to the CEFR; and it is less prone to statistical biases (Reckase, 2006). The use of additional standard-setting methods, such as the Basket Method ,would provide multiple sources of evidence, in keeping with the principles above. Following Hurts and Hertz (1999), panels of 10 to 15 judges were organised for each standard-setting exercise. Training and training materials were provided to ensure familiarity with the CSE, and a number of instruments were also used to capture evidence related to the procedural validity, internal validity, and external validity of the exercise.

Stage 3 – External Validation: External validation was carried out through the Contrasting Groups method by gathering actual test data from students in addition to teacher judgements to place students in a particular CSE level. Cizek and Bunch (2007) describe a number of different methods for determining cutscores using the Contrasting Groups method.

In addition, reference has been made to the relationship between the tests and the CEFR, and between the levels of the CSE and the CEFR. The CSE development project incorporated the CEFR descriptors into the scaling of the CSE descriptors, and work has been done on mapping CEFR levels against the CSE. Thus, it is possible to triangulate data from several studies linking the tests to the CEFR and information collected during the development of the CSE on how CEFR levels might be mapped to the CSE levels.

The descriptions which follow focus on the triangulation of evidence across the three distinct stages.

5.3.3 Stage 1: Construct evaluation

In order to investigate the construct targeted by the exams, a model of language test validation is essential. For this study, the socio-cognitive model of language test development and validation was employed to underpin the overarching design and evaluation of data collection in relation to the constructs targeted by the tests. More information on the socio-cognitive model itself is provided in Section 5.4. Here, those elements of the model are introduced which make it particularly amenable to adaptation for this key stage in the linking framework.

The socio-cognitive model has built on advancements in validity theory since the 1990s to create a model which provides an explicit and comprehensive framework, with taxonomies of features relevant for the different skill areas, designed for describing both the contextual features and the cognitive processing demands of test tasks. It is the latter feature in particular which has made the socio-cognitive model a particularly powerful tool for describing and validating, the profile of features relevant to test tasks targeting different levels of proficiency. The model was first fully presented as a set of frameworks across all four skills by Weir (2005), and was further updated by O'Sullivan and Weir (2011). O'Sullivan (2011, 2015a, 2016) has made further modifications to the model.

A number of comprehensive case studies demonstrating how the model has been applied in practice have been developed, and these have helped to refine and add to the taxonomies of features described, including Cambridge examinations (Geranpayeh & Taylor, 2013; Khalifa & Weir, 2009; Shaw & Weir, 2007; Weir, Vidakovic & Galaczi, 2013), the Aptis test system (O'Sullivan, 2015a; O'Sullivan & Dunlea, 2015), the TEAP test in Japan (Nakatsuhara, 2014; Taylor, 2014; Weir, 2014), the GEPT test in Taiwan (Wu, 2014), and the EIKEN test in Japan (Dunlea, 2016). The model has further provided the descriptive framework to facilitate the interpretation of evidence gathered through a range of mixed-method studies involving think-aloud protocols, questionnaires, and eye-tracking, particularly in relation to both IELTS and Aptis (e.g., Bax, 2013; Brunfaut & McCray, 2015).

The taxonomies of contextual parameters used to describe test tasks in these studies have also drawn heavily on the grids for describing reading and listening tasks first developed by Alderson et al. (2006) and now included in the *Manual for Linking Exams to the CEFR* (Council of Europe, 2009). These strong connections across studies, and particularly in relation to linking to the CEFR, have built up a powerful body of evidence that allows for comparison of examinations claiming to target particular skill areas and particular proficiency levels. These claims can thus be evaluated by making reference to these shared sets of contextual and cognitive parameters provided by the socio-cognitive model.

Wu (2014) and Dunlea (2015) have drawn on this extensive body of studies to develop and refine evaluation templates which can be used to construct profiles of the contextual and cognitive parameters of test components. These evaluation instruments have been further refined in two test comparability studies. The first study explored the relationship between the GEPT and Aptis, using the claims of each test's relevance to the CEFR as a central point of comparison (Wu et al., 2016). Refining the methodology in this study, Dunlea et al. (2018) further developed the proformas to develop detailed task profiles across contextual and cognitive parameters to help compare the constructs targeted by Aptis and VSTEP, a national standardised test of proficiency developed in Vietnam. In the latter study, the evaluation templates were applied by two groups of trained raters, one in Europe and one in Vietnam, for the purposes of building up task profiles. Feedback on the usefulness of the evaluation templates was obtained from both groups to further enhance the instruments for use in future comparability studies.

For this particular CSE linking study, the evaluation templates developed for use in the Aptis–VSTEP study were adapted for use in building a comprehensive set of task profiles for the test tasks which make up the examinations in this study. The categories included are grounded in the socio-cognitive model and have been derived from the extensive body of research described above, as well as having clear overlap with the grids used in the Manual (2009). At the same time, the categories have been drawn directly from publicly available test specifications used, for example, in Aptis (O'Sullivan & Dunlea, 2015) and TEAP (Taylor, 2014), and so have clear practical application for actual test description. The templates were then further modified to include judgements of which CSE descriptors were relevant on a task-by-task basis. The templates include an estimation of the appropriate CEFR level for each task in the test being evaluated. While based on expert judgement, the templates thus also provide a method of triangulating proficiency estimations across the two frameworks (CSE and CEFR) within the same template.

To carry out this phase, a sub-group of the Working Group, consisting of two researchers each from Cambridge Assessment and the British Council, coordinated the evaluation. The researchers met first to review the templates and ensure that they had a consistent interpretation. The Cambridge Assessment researchers then analysed the Aptis test across all four skills, mapping specific CSE descriptors to each task. The British Council researchers carried out the same process for IELTS. In each case, the researchers first worked independently to make judgements using the template and any discrepancies were discussed and resolved resulting in a consensus version of evaluation templates filled in for each test.

5.3.4 Stage 2: Test-centred (expert panel) standard setting

5.3.4.1 Overview

As described in the literature above on standard setting, in particular in relation to linking of exams to descriptive frameworks of proficiency such as the CEFR or CSE, standard setting plays a crucial role in gathering evidence to support the claims of test developers that a particular exam can be considered relevant to, and able to measure, a test-taker's proficiency in relation to such frameworks. As already described, one broad distinction often employed in standard setting is between test-centred (expert panel) approaches and examinee-centred (data driven) approaches. In practice, test-centred approaches are employed the most. Although relatively intensive in terms of the demands placed on organisers and facilitators of such panels, they have the practical advantage of being focused on a small number of participants and centred on a limited timeframe. Expert panels typically take no more than several days to set cutoff points for a single exam. However, they are logistically demanding in that panels of experts need to be recruited who have the required background and suitable availability. In fact, as described in the literature review, an important part of validating any standard-setting study is to gather procedural validity evidence from the participants through questionnaires and other procedures.

5.3.4.2 The expert panel approach

As the literature above indicates, a group of 10 to 15 judges is more than sufficient to meet the standards of best practice in the field. For the purposes of this study, and given the nature of the exams, panellists were drawn largely from educators working in higher education in China with a good understanding of the local context. Since in the initial stages of the study, the CSE had yet to be released and was still relatively unfamiliar to participants, the composition of the expert panel was determined such that panel members would be familiar with a broad range of key concepts and points in the CSE, drawing on accumulated experience and relevant curriculum documents, and would have an understanding of levels of performance relevant to particular educational sectors and levels etc. Familiarity with this context facilitated the training process, in that it permitted greater focus on the standard-setting methodology and understanding of the exams involved. All participants were required to have a high level of English to be able to evaluate the test items, which range up to C1.

5.3.4.3 Training for the panel

Standard setting in relation to frameworks such as the CEFR or CSE normally includes training and discussion of the descriptions of performance contained within the calibrated descriptors which make up the framework. However, building a shared understanding of these descriptors is still a necessarily judgemental and somewhat subjective endeavour. In standard setting in relation to the CEFR, best practice usually involves also training and practice at estimating the level of test items and tasks which have already been calibrated to the framework and which can be considered to some extent to be concrete operationalisations of the verbal descriptions in the descriptors (or performance level descriptors, as they are known in standard setting). However, given that the CSE was only publicly released during the span of the linking project, no tests had yet been developed from the CSE. This element of training was thus not possible to implement for this project.

Time constraints on the availability of panel participants are a common hurdle for standard-setting panels across contexts. Dunlea (2015) overcame this in a series of extensive standard-setting studies in Japan by employing self-access study guides to help participants prepare for the standard-setting panels. Members of the Working Group who had also belonged to the original CSE development project took the lead in producing a comprehensive self-study booklet which contained detailed descriptions of the CSE and familiarisation activities based on those found in the *Manual for Linking Examinations to the CEFR* (Council of Europe, 2009). These activities were adapted to provide participants with hands-on experience in manipulating the CSE descriptors, for example, reordering sets of jumbled descriptors into the right proficiency levels, identifying criterial features which distinguish one level from another, etc.

The self-access materials can provide an important foundation so that training during the panel sessions can move forward much more quickly. For each standard-setting panel, the first day of the actual face-to-face meetings was allocated to concentrating on training with the CSE descriptors for the particular skill that was the focus of linking (reading, listening, speaking or writing). These sessions were led by researchers from the Working Group with extensive experience in facilitating standard-setting panels. Panellists reviewed the activities they had carried out in the self-access study guide, and engaged in discussion to identify key criterial features of particular CSE levels. Finally, a consensus version of key words and defining features that distinguish particular CSE levels was produced on screen. While the consensus list of criterial features helped focus attention in the group on an agreed set of key aspects for each CSE level, the facilitators stressed to the participants the importance of returning to the full list of CSE descriptors and scales while making judgements, and not to rely only on the shorthand summary list of key features produced during the training. Several versions of the CSE descriptors were produced for the panellists. In one, the descriptors were collated according to each separate subscale, progressing from the lowest level to the highest level within that subscale. Another version was produced in which all descriptors were collated according to level, with all descriptors from all subscales being presented within each level. Panellists were encouraged to use whichever version they found most convenient, and of course had the consensus summary list of key features to refer to as well. The working language of the panels was English. All descriptors had been translated into English prior to the panels.

5.3.4.4 The test-centred standard setting methods

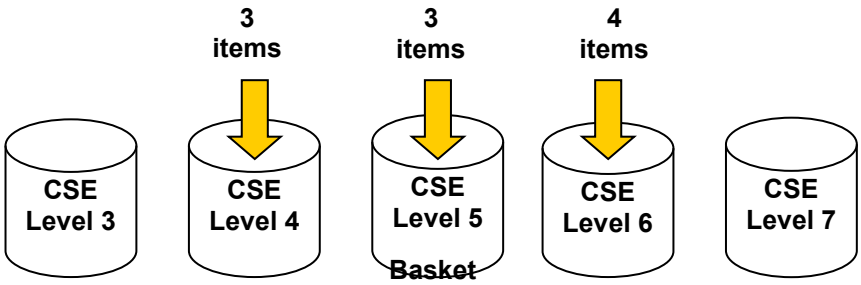
This project employed two test-centred standard-setting methods for the receptive skills. This approach and the same two methods were employed by Dunlea (2015) and O'Sullivan (2015b) in relation to the EIKEN and Aptis exams, respectively. Dunlea (internal technical report) has also employed a similar approach successfully with a later variant of Aptis, Aptis Advanced. This approach involves the same expert panel being trained in, and making judgements with, two test-centred methods, the Basket Method and a Modified Angoff Method.

Based on the literature review above, it is clear that although some misgivings have been voiced regarding the Modified Angoff Method, it remains one of the most widely used methods both in relation to standard setting related to the CEFR and also in the field of educational measurement in the United States itself. A number of studies have, in fact, refuted the claims that it is cognitively too demanding, particularly for teachers accustomed to considering standardised test results for their students, as is the case in China. In addition, the studies by Reckase cited above also underscore that the Modified Angoff Method is less prone to statistical bias than other popular methods such as the Bookmark Method. When utilising a Modified Angoff approach, experience has shown (Dunlea, 2015; O'Sullivan, 2015a) that first engaging with a more intuitively simple method can help participants, particularly when they are trying to also conceptualise a very new framework as the performance level descriptions. As such, the project replicated the approach used in Dunlea (2015) and O'Sullivan (2015b) by using the Basket Method to first help participants form broad-brush estimates of how particular items relate to particular levels of the CSE. Following this initial stage, several rounds of the more complex Angoff Method were carried out, with these more statistically robust rounds of data collection being used to estimate actual cutoffs.

5.3.4.5 Basket and Modified Angoff Method procedures for Listening and Reading

The application of the Basket Method employed for this project involves deciding in which of several "baskets" particular test items belong. To allocate an item to a basket, a panellist considers at what level in the framework in question a learner would first be able to complete the item successfully. A diagrammatic representation of the idea behind the method can be seen in Figure 3

Figure 3: Example of basket method item allocation



In this case, the baskets are in fact the CSE levels. If one were to use the Basket Method to set cutoffs, the procedure, in its simplest form, would be derived by adding up the total number of items in all of the baskets below the level in question and adding one. For example, in the example in the diagram above, the test has 10 items, and three were placed in the CSE Level 4 basket, three in the CSE Level 5 basket, and the remaining four in the CSE Level 6 basket. In its simplest form, the cutscore for a test-taker to be considered at CSE Level 6 would be the number of items in those baskets below the Level 6 basket plus 1. In this case, there are a total of six in the Level 4 and Level 5 baskets, so a test-taker would need to get seven or more items correct to be considered at CSE Level 6.

However, it is important to note that this project, from the outset, intended to employ the Basket Method as an "ice breaker" to help panellists become familiar with the items in the test and consider those items in terms of key features of the CSE. This activity allows panellists to establish an initial hypothesis regarding the approximate area of relevance for the item in terms of CSE levels. When revisiting items to make the more cognitively demanding probability judgements per item required by the Modified Angoff Method, panellists can usefully refer back to their Basket Method judgements to establish an approximate probability range, refining that judgement to arrive at a final judgement for the item. This range-finding process expedites and streamlines the standard-setting procedure, allowing panellists to home in on relevant levels more quickly. As such, the Basket Method was not intended to be used for setting cutscores to determine the boundaries between CSE levels on the tests in question.

For both the Listening and Reading components of Aptis and IELTS, the following judgement task, or question, was posed to the panellists when using the Basket Method: *Review each item: at which (CSE) level can a minimally competent examinee FIRST answer this item correctly?*

For the Basket Method, panellists first carried out one round of judgements. Panellists were asked to take the test first – answering questions as would test-takers – before making judgements about each item. After initially answering the items, panellists were presented with the item key. Judgements were entered into specially prepared Excel rating forms individually by each panellist. After all panellists had entered their judgements, the Working Group project members collected the judgements and collated results as a series of bar graphs, one for each item, to display to panellists the distribution of ratings across CSE levels for each item. This was used by the facilitators to elicit discussion, focusing on items that had a greater degree of variability in level judgements. Panellists were encouraged to offer the rationale for their ratings. It was stressed to panellists that this normative feedback was for reference only and that no panellist was required to modify their ratings based on this feedback. Following discussion, panellists were offered the chance to change any ratings they wished to modify, and the final version of their ratings were collected and collated by the project team

The Modified Angoff Method employed involves making probability judgements about each item. Using a population variation of the judgement task recommended by Cizek and Bunch (2007), in this project panellists were asking to imagine an ideal group of 100 candidates who are minimally competent at a particular level, for example CSE Level 6. The procedure then calls for judges to estimate the number of test-takers in this hypothetical group of 100 who would get each item correct.

In the example above, where the first item was judged to be Level 4, a judge might estimate that a large number of this group of 100 CSE Level 6 test-takers will correctly answer the item. For one of the more difficult items placed in the Level 6 basket above, the judge might estimate that a smaller number, perhaps 50 out of 100, or 50% would answer the item correctly. Judges repeat this process for all items in the test. Cutscores can be set by calculating the mean of the percentage correct estimates across all raters for each item, then calculating the mean of these mean estimates across all items. In addition to this method, in this study, a Multi-facet Rasch Model using the FACETS program (Linacre, 2014) was employed to calculate a fair average of the probability judgements for each item. The mean of the fair average estimates from the second round of judgements was used to estimate the cutscore for each level.

The judgement task for both Listening and Reading for the Modified Angoff Method was put to the panellists in the following way: *Imagine 100 candidates at the same level. Examine each item. How many minimally competent examinees at the target level will answer the item correctly? Only use 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100.* Panellists first carried out this task for the lowest CSE level in the target range for all items. This judgement process was then repeated for each CSE level in the target range (i.e. three to seven for Aptis and four to eight for IELTS).

With the Angoff Method, panellists were asked to go back to the beginning of the test, reviewing each item once again as a test-taker before making their judgements. For the Listening test, this meant asking all panellists to listen to the full test again from start to finish. Two full rounds of judgements were conducted. As with the Basket Method, after the first round of judgements were completed, the project team collected results and collated them into descriptive tables with the mean, median, mode, maximum and minimum percentage correct estimates for each level for each item, as well as graphically displaying the results. As with the Basket Method, panellists were encouraged to provide their rationale for their ratings. Discussion once again focused mainly on items showing wide dispersion of results to shed light on the reasons for the differing interpretations. Following this, panellists were then given empirical feedback in the form of the facility values for each item from a live administration of the test (for Aptis, this was a large-scale field trial carried out as a part of the revision process). Care was taken to explain to panellists that this empirical feedback needed to be treated carefully and probability judgements should not be adjusted to simply reflect the facility values. This is because the test results from live administrations contained test-takers from a wide range of proficiency levels, whereas the judgement process focuses on a hypothetical example of 100 minimally competent candidates at the same target level. Nonetheless, the empirical feedback was offered as a useful way of identifying items which were significantly more or less difficult than a panellist may have anticipated and was once again used for reflection and discussion. After the normative feedback, empirical feedback, and discussion, a full second round of judgements was carried out. Panellists once again reviewed all of the items individually and input ratings into a second-round judgement form prepared for the purpose. These results were once again collected by the project team for post-hoc analysis.

5.3.4.6 The examinee-centred method used in Speaking and Writing panels

5.3.4.6.1 Overview

The methodology for both speaking and writing was a modified Analytic Judgement Method (AJM) (Cizek & Bunch, 2007; Plake & Hambleton, 2000; Hambleton et al, 2000). The AJM procedure is an examinee-centred approach according to Jaeger's (1989) classification. The AJM procedure uses examples of actual student performance on a test. The AJM procedure as originally described selects a student performance based on their total score performance on the test, but then breaks those performances up into major components within the test. Panellists address each major component separately, allocating student samples within that component to the major performance categories for which standards need to be set. Within each component, however, samples cover multiple questions or tasks. In this respect, the AJM shares features of holistic, body of work procedures (Cizek & Bunch, 2007) in that each performance sample for a component comprises a collated body of work across all questions / tasks in that component for an examinee. In the original AJM, cutoffs would be set for each component, and the sum of cutoffs across components would be the cutoffs for a particular performance standard or level on the whole test. These methods share many aspects also with the Contrasting Groups / Borderline candidate examinee-centred methods.

While typical Contrasting Group methods call for students to be allocated to levels by teachers familiar with the students (Cizek & Bunch, 2007), adaptations have been applied in which judgements are made of student performance samples by judges not familiar with the students (e.g. Bechger, Kujper & Maris, 2009). Recently Knoch and Frost (2016) applied a modified Contrasting Groups approach to linking the GEPT writing tests to the CEFR which, in practice, could be said to combine aspects of both the AJM and Body-of-Work approaches with traditional Contrasting Groups / Borderline Group methods for setting standards.

Our approach needed to take into consideration practicality in that two different skills (speaking and writing) needed to be addressed for two different tests (Aptis and IELTS) within a limited timeframe of five days. The method adopted was thus an adapted AJM procedure. While the AJM procedure as described in Plake and Hambleton (2000) and Hambleton et al. (2000) was applied to setting a composite cutoff for each proficiency level across multiple components, the panels for Speaking and Writing in this project treated a full test performance for each skill (speaking or writing) for each test (Aptis or IELTS) as consisting of only one component. The collated performance samples across all tasks within one full test were thus addressed as one performance sample, with one level estimate being provided for the entire test performance sample.

Below, we first give an overview of the features of the method common to both Speaking and Writing, before addressing features particular to each skill.

Samples of performance for each skill for each test were selected based on the overall or total score so as to represent all possible score points. Panellists formed a consensus interpretation of the CSE levels for each skill based on the preparation booklet and the face-to-face training during the event. Following Plake and Hambleton, panellists used a direct classification approach to allocate the performance samples to the relevant CSE levels. However, each target level was further broken down into low, mid, and high categories. For Aptis, this resulted in the following classification categories: below CSE 2, CSE 3 low, CSE 3 mid, CSE 3 high, CSE 4 low, CSE 4 mid, CSE 5 low, etc. through the target levels for Aptis, and finishing with CSE 8 as the highest levels. For IELTS, the categories had below CSE 3, CSE 3, CSE 4 low, CSE 4 mid, CSE 4 high, etc. through the target levels for IELTS with the highest category being CSE 9. This method allows for two alternative analysis procedures. Ideally, this would involve combining the high and low categories of adjacent levels into a borderline category, with the cutoff being set through several possible procedures, including the mean of the original ratings (Aptis scale scores, IELTS band score) or several alternative methods. In the event that the number of performances allocated to the borderline categories is too low, sub-categories within a level can be collapsed to increase the number of ratings and calculation procedures similar to the Contrasting Groups procedures described in Cizek and Bunch (2007) can be used as an alternative for estimating cutoffs.

For both Speaking and Writing, two rounds of judgements were carried out, with presentation of normative feedback and discussion between rounds. No empirical feedback was given. In this method, panellists are required to allocate the performances to CSE levels without knowing the original score (Aptis scale score or IELTS band score).

5.3.4.6.2 Speaking

Speaking is particularly constrained by practical realities such as the time available, as panellists are required to listen to an entire performance. To maximise the number of performances rated for analysis purposes, panellists were split into four groups. All groups rated a common set of 10 performances for each test. Following this, each group separately rated a set of 15 performances unique to that group. This resulted in a set of 70 performances in total. While only four ratings were collected for the performance samples unique to each group, using MRFM analysis with the FACETS program for the common linking set of 10 performances, we were able to place all performances on a common scale. This allowed us to use the fair average for that performance as the final CSE level it was assigned to. For Speaking, the first round rating results for the common linking set of 10 Speaking test performances were used for discussion before proceeding to the second round of judgements.

5.3.4.6.3 Writing

The rating of writing performances proceeded more quickly than speaking performances, and as such all raters were able to rate the same batch of 120 writing test performances, providing a fully crossed design. Raters, however, remained in the four groups they were divided into for Speaking. The first 100 performances were placed in the same order for all raters. The remaining 20 performances were ordered differently for each group. This was to ensure that if any problems arose and time for rating became limited, ratings would still be collected on all 120 samples ensuring a very robust set of performances for Writing. As with Speaking, a first round of judgements was carried out and collected. Results were collated during lunch, and a set of performances with noticeable variability was selected for presentation and discussion. Following discussion, raters carried out a full second round of judgements on the same 120 Writing test performances.

5.3.4.7 Stage 3: Examinee-centred (data driven) standard setting

As described above in the literature review, Kane has emphasised that employing both test-centred (expert panel) and examinee-centred (test data driven) approaches is a potentially powerful form of validating cutscores. Dunlea (2015) has demonstrated how these two approaches can, in fact, be combined and employed to provide a powerful form of external validity for the standard-setting process and the claim of a link. Following the methodology employed by Dunlea (2015), and described also in Dunlea and Figueras (2012), the test-data driven part of the study employed a Contrasting Groups standard-setting methodology. Various applications of the Contrasting Groups Method are described in the literature (e.g. Green, Trimble & Lewis, 2003; Livingston & Zieky, 1989; Van Nijlen & Jansenn, 2008; and Bechger, Kujper & Maris, 2009). The method used in Dunlea (2015) drew on these descriptions, in particular the interesting adaptation of the method in relation to the CEFR described by Bechger, Kujper and Maris (2009). The method involved actually administering both tests to a large group of students. At the same time, teachers familiar with the students in question were trained in the relevant areas and levels of the CSE for the study. Teachers made judgements about their students, allocating students to the levels of the CSE that they felt best reflected the students' levels of proficiency. Using the actual test results of the students, score distributions for each of the groups of students allocated to each CSE level could be calculated. A number of methods are then available for estimating the cutoff points, as described in Cizek and Bunch (2007).

The sampling methodology and recruitment of participants, as well as training of teachers taking part, is described in more detail in Section 7, under the results for the external validation part of the project.

5.4 The socio-cognitive model

The CSE, as described in Section 4, has drawn on a comprehensive range of theories modelling language proficiency to inform the development of the descriptors and structure the scales. We will here briefly introduce another model of language test development and validation, the socio-cognitive model as this model has been used extensively in the design and validation of the exams that are the focus of our linking activities, and has provided the theoretical framework for the methodology underpinning the linking project.

The socio-cognitive model for language test development and validation was first fully elaborated, with validation frameworks describing criterial features across each of the four skills, in Weir (2005), and has been elaborated and developed further in O'Sullivan (2011, 2015a, 2016) and O'Sullivan and Weir (2011). In its initial formulation in Weir (2005), the model contained five aspects of validity essential for collecting evidence which would support a balanced, coherent, and comprehensive validity argument in support of the uses and interpretations of a test: content validity evidence, cognitive validity evidence, scoring validity evidence, criterion-related validity evidence and consequences and impact validity evidence. Dunlea (2015) has shown how these five categories of evidence overlap with the six aspects of validity evidence which Messick suggested were necessary, and sufficient, to ensure that "the theoretical rationale or persuasive argument linking the evidence to the inferences drawn touches the important bases" in a comprehensive validity argument to justify the uses and interpretations of a test (Messick, 1995, p. 747). It is worth noting in advance that the model does not try to replace earlier important milestones in validity theory, such as Messick's seminal 1989 encapsulation of the unitary concept of validity, or indeed, the important contribution of Bachman's CLA model.

Rather, it attempts to incorporate the lessons learned since those milestones while building on the foundations and important principles which have become accepted as underpinning the field. A large body of literature has now been established utilising the socio-cognitive model in both retrospective validation studies (Dunlea, 2015; Khalifa & Weir, 2009; O'Sullivan, 2010; Shaw & Weir, 2007; Taylor, 2012; Geranpayeh & Taylor, 2013; Wu, 2014) and also to drive test design and development from the outset for new tests, (Nakatsuhara, 2014; O'Sullivan, 2015a; O'Sullivan & Dunlea, 2015; Taylor, 2014; Weir, 2014).

As already noted, the model has been employed in the design of Aptis from the outset and extensively in research into the IELTS test; from this perspective alone, it is worth taking into consideration as we attempt to understand the two examinations and the way they measure language proficiency. In addition, in initial discussions within the Steering Group to evaluate appropriate methodology for the linking project, it was noted that two particular features of the model interact with important design features of the CSE.

The first is the explicit incorporation of the cognitive demands posed by test tasks into the validity evidence framework, which has been a major contribution of the model. This complements the important work of the CSE in attempting to make explicit the role of cognitive processing in the descriptor design, something which remained implicit within the CEFR. In validating a test, it is essential to establish the cognitive profile of the test tasks, elaborating the cognitive processes that will be elicited when test-takers engage in a task. This then enables validation through a comparison of whether these processes resemble similar processes elicited when language users engage in real-life language use tasks in the TLU. This has led to the inclusion of intended task cognitive profiles in task specifications where the model has been used to drive test design (see for example O'Sullivan & Dunlea, 2015; Taylor, 2014; Nakatsuhara, 2014; Weir, 2014). The socio-cognitive approach has resulted in important developments in designing models of cognitive processing and building these processes into test design in a way that is amenable to clear specification and consequently validation and empirical verification. (e.g. Bax, 2013; Bax & Weir, 2012; Brunfaut & McCray, 2015).

The second important area that the socio-cognitive model intersects with the CSE is the explicit recognition of the importance of the social context of use in which tests are embedded. Just as cognitive processing profiles are an important part of test specification, the model has also emphasised building detailed taxonomies of contextual features useful for describing test tasks. These contextual features will interact with cognitive features, and can be manipulated to derive language test tasks targeted at different levels of difficulty. This is something obviously of importance to an educational framework such as the CSE, used within a TLU which, as noted in Section 5, is often embedded in an educational EFL context for many Chinese learners. The social aspect of the socio-cognitive model also extends to another level outside the specification of test features and how the test-takers interact with these features in taking the test itself. Another important dimension is the way a test system is embedded within a social context of use, and the model aims to tease out and make explicit the relationship between stakeholders, test system design, test use, and the wider social context of that use. Recent iterations of the model in O'Sullivan (2015a, 2016) have reorganised the visual presentation of evidence under three core areas of the test-taker, the test system (which includes descriptions of contextual and cognitive features of test tasks) and the scoring system, which O'Sullivan and Weir (2011) suggest could usefully subsume criterion-related evidence rather than having this as a stand-alone category. O'Sullivan (2016) has further embedded these features within the social context of use, visually representing the ongoing interaction between key stakeholders and a test system, emphasising that impact flows in both directions, not just in a linear fashion from test to teachers and learners. This has important implications for the linking project, and can provide useful ways of evaluating key design decisions of the CSE itself within the socio-cognitive framework.

The CSE, as noted above in Sections 5 and 6, has clearly been designed within and for a specific context of use, and key design decisions are posited as being made in response to the particular demands of that context. The socio-cognitive model will thus provide a useful theoretical framework for collating, evaluating and presenting the validity argument around this linking project itself, and for describing how the features of the tests under consideration relate to features of the CSE relevant to Chinese learners.

6. DESCRIPTION OF THE TESTS

6.1 IELTS

IELTS is an international test of English proficiency testing all four skills. IELTS is jointly owned by the British Council, IDP: IELTS Australia and Cambridge Assessment English, and is used extensively around the globe for a number of purposes including university entrance.

There are two types of the IELTS test: IELTS Academic and IELTS General Training. The Listening and Speaking papers are the same for both tests, but the subject matter of the Reading and Writing components differs depending on which test is taken. For the standard-setting activities, the IELTS Academic Reading and Writing tests were used.

The Listening, Reading and Writing components of all IELTS tests are completed on the same day, with no breaks in between them. The Speaking component, however, can be completed up to a week before or after the other tests. The total test time is 2 hours and 45 minutes. The structure of the IELTS Listening, Academic Writing, Academic Reading and Speaking papers are outlined in the tables below.

Table 2: IELTS Listening test description

Paper format	There are four sections with 10 questions each. The questions are designed so that the answers appear in the order they are heard in the audio. The first two sections deal with situations set in everyday social contexts. In Section 1, there is a conversation between two speakers (for example, a conversation about travel arrangements), and in Section 2, there is a monologue in (for example, a speech about local facilities). The final two sections deal with situations set in educational and training contexts. In Section 3, there is a conversation between two main speakers (for example, two university students in discussion, perhaps guided by a tutor), and in Section 4, there is a monologue on an academic subject. The recordings are heard only once. They include a range of accents, including British, Australian, New Zealand, American and Canadian.
Timing	Approximately 30 minutes (plus 10 minutes transfer time).
No. of questions	40
Task types	A variety of question types are used, chosen from the following: multiple choice, matching, plan/map/diagram labelling, form/note/table/flow-chart/summary completion, sentence completion.
Answering	Test-takers write their answers on the question paper as they listen and, at the end of the test, they are given 10 minutes to transfer their answers to an answer sheet. Care should be taken when writing answers on the answer sheet as poor spelling and grammar are penalised.
Marks	Each question is worth 1 mark.

Table 3: IELTS Academic Reading test description

Paper format	Three reading passages with a variety of questions using a number of task types.
Timing	60 minutes
No. of questions	40
Task types	A variety of question types are used, chosen from the following; multiple choice, identifying information, identifying the writer's views/claims, matching information, matching headings, matching features, matching sentence endings, sentence completion, summary completion, note completion, table completion, flow-chart completion, diagram label completion and short-answer questions.
Sources	Texts are taken from books, journals, magazines and newspapers, and have been written for a non-specialist audience. All the topics are of general interest. They deal with issues which are interesting, recognisably appropriate and accessible to test-takers entering undergraduate or postgraduate courses or seeking professional registration. The passages may be written in a variety of styles, for example, narrative, descriptive or discursive/argumentative. At least one text contains detailed logical argument. Texts may contain non-verbal materials such as diagrams, graphs or illustrations. If texts contain technical terms, a simple glossary is provided.
Answering	Test-takers are required to transfer their answers to an answer sheet during the time allowed for the test. No extra time is allowed for transfer. Care should be taken when writing answers on the answer sheet as poor spelling and grammar are penalised.
Marks	Each question is worth 1 mark.

Table 4: IELTS Academic Writing test description

Paper format	Two task types requiring one shorter and one longer written piece.
Timing	60 minutes (20 minutes for Task 1 and 40 minutes for Task 2 approximately)
No. of questions	2
Task types	Task 1 – candidates are shown a graph, table, chart or diagram and asked to describe, summarise or explain the information in their own words. Describe and explain data, describe the stages of a process, how something works or describe an object or event. Task 2 – candidates write an essay in response to a point of view, argument or problem. Responses to both tasks must be in a formal style.
Sources	Topics are of general interest to, and suitable for, test-takers entering undergraduate and postgraduate studies or seeking professional registration
Answering	Candidates must handwrite a response to each task (1 and 2).
Marks	Each task is marked (1–9) according to band score descriptors, and an overall grade is given. Categories include: Task Achievement, Coherence and Cohesion, Lexical Resource, Grammatical Range and Accuracy.

Table 5: IELTS Academic Speaking test description

Paper format	Three parts including an initial exchange, monologue long turn then follow-up discussion.
Timing	11–14 minutes (4–5 minutes for Part 1, 4 minutes including preparation for Part 2, 4–5 minutes for Part 3).
No. of questions	3
Task types	Part 1 – the examiner asks general questions about yourself and a range of familiar topics, such as home, family, work, studies and interests. Part 2 – the candidate is given a card which asks you to talk about a particular topic, then has one minute to prepare before speaking for up to two minutes. The examiner will then ask one or two questions on the same topic. Part 3 – the candidate is asked further questions about the topic in Part 2. These provide the opportunity to discuss more abstract ideas and issues.
Sources	Common topics relate to personal experience and daily life, such as hometown, studies, work, free time, family and a variety of other subjects designed to encourage spoken communication.
Answering	Candidates' spoken monologue and subsequent exchange are recorded.
Marks	Candidates are given an overall score (1–9) according to band score descriptors. Categories include: Fluency and Coherence, Lexical Resource, Grammatical Range and Accuracy, Pronunciation.

More detailed information on the format of the IELTS Academic Reading, Writing, Speaking and Listening parts of the test is available online at: https://www.ielts.org/about-the-test/test-format-in-detail

Sample academic reading, writing, speaking and listening items can be tried at: https://www.ielts.org/about-the-test/sample-test-questions

6.2 Aptis

The Aptis test system is an approach to test design and development devised by the British Council. Tests are developed within the Aptis system for various uses by different test users. Aptis General, the main variant within the system, is a test of general English proficiency for adult test-takers. It is offered directly to institutions and organisations for testing the language proficiency of employees, students, etc. Aptis General is designed to provide assessment options for ESL/EFL speakers spanning proficiency ranges from A1 to C1 in terms of the CEFR. Test-takers will be 16 years old or older. Learners may be engaged in education, training, employment or other activities. Aptis has five components, targeting all four skills with an additional Core component targeting grammar and vocabulary. Aptis is a computer-based test, but all components can also be taken as a pen-and-paper test.

Descriptions of the Aptis General Listening, Reading, Speaking and Writing components are provided in the tables below. Note the Aptis General test has been the subject of a revision project and that the descriptions below are revised versions of the test which are not yet publicly available at the time of writing of this report. These revised test components were used during the standard-setting meetings.

More detailed information about the current Aptis test live at the time of writing, which includes sample materials and is aimed at a non-specialist audience, can be accessed here: https://www.britishcouncil.org/exam/aptis.

A full description of the test system live at the time of writing, including detailed task-level specifications, is provided in the *Aptis General Technical Manual* (O'Sullivan and Dunlea, 2015) at https://www.britishcouncil.org/aptis-general-technical-manual-version-10

Table 6: Aptis Listening test description

Skill focus	Items/ level	Format	Task description	Response format
Lexical recognition	5	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	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	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 drag and drop. 6 written options. Distractors should have some overlap with information and ideas in the texts.
Meaning representation/ inference	4	Dialogue	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.	2x3-option multiple choice. Both target and distractors are paraphrased or implied, and distractors refer to information and concepts in the text that are not possible answers to the question

Table 7: Aptis Reading test description

Skill focus	Items/ level	Task focus	Task description	Response format
Sentence level meaning	5	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	3	Inter-sentence cohesion (Careful global reading)	Reorder 5 jumbled sentences to form a cohesive text	Text consisting of 6 sentences. The first sentence is fixed. Candidates reorder the following 5 sentences.
	3	Inter-sentence cohesion (Careful global reading)	Reorder 5 jumbled sentences to form a cohesive text	Text consisting of 6 sentences. The first sentence is fixed. Candidates reorder the following 5 sentences.
Text-level comprehension of short texts	7	Text-level comprehension of short texts (Global reading, both careful and expeditious)	Matching 7 statements of opinion with people associated with different texts. Selecting the correct person requires text-level comprehension and reading beyond the sentence containing the gap.	4 short paragraphs. Candidates choose from a drop-down menu, which of the four people could say certain statements.
Text-level comprehension of long text	7	Text-level comprehension of longer text (Global reading, both careful and expeditious)	Matching the most appropriate headings to 7 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 8: Aptis Speaking test description

Part	Skill focus	Task description	Channel of input / prompts	Time for response
1	Giving personal information.	Candidate responds to 3 questions on personal topics. 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).	30 seconds to respond to each question. No planning time.
2	Describing, expressing opinions, providing reasons and explanations.	The candidate responds to 3 questions. The first asks the candidate to describe a photograph. Followed by 2 questions 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.	45 seconds to respond to each question. No planning time.
3	Describing, comparing and contrasting, providing reasons and explanations.	The candidate responds to 3 questions / prompts and is asked to describe, contrast and compare 2 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) 2 photographs showing different aspects of a topic are presented on screen.	45 seconds to respond to each question. No planning time.
4	Integrating ideas on an abstract topic into a long turn. Giving & justifying opinions, advantages and disadvantages.	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 2 minutes to present a coherent, continuous, long turn.	1) 3 questions are presented simultaneously in both written and oral form (pre-recorded). Questions remain on screen throughout the task. 2) 1 photograph illustrating an element of the topic mentioned in the prompts. The photo is not referred to in the questions.	2 minutes. Responses to the 3 questions are integrated into a single long turn. 1 minute planning time.

Table 9: Aptis Writing test description

Part	Skill focus	Task description	Channel of input / prompts	Expected output
1	Writing at the word or phrase level. Information to simple questions in a text message type genre.	The candidate answers simple questions. All responses are at the word or phrase-level. Each response will consist of responses to 5 questions.	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.
2	Short written description of concrete, personal information at the sentence level.	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.	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.	The candidate writes 2 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.

7. RESULTS

7.1 Construct definition

As mentioned above, the construct definition stage was carried out by a subgroup of the Working Group consisting of two researchers each from Cambridge Assessment and the British Council. The Cambridge Assessment researchers analysed the Aptis test across all four skills, while the British Council researchers did the same for IELTS. In each case, the researchers completed the construct definition templates generated in accordance with the socio-cognitive model (Appendix C), mapping CSE descriptors to each task and agreeing upon a consensus version for each test. The judges identified the key contextual and cognitive parameters and allocated CSE descriptors to each task on the basis of their own interpretation of the task features, rather than any documentation produced by the relevant test developer. In a final cross-validation stage, these construct definitions were then exchanged between the two teams of researchers with the result that consensus was reached and the final judgements were made with a high degree of confidence.

It is important to observe that more CSE descriptors exist at the lower levels (e.g. 3–5) than at higher levels (e.g. 7–9). Therefore, in the summary of results that follows, there may be a tendency for lower level descriptors to have greater representation.

7.1.1 IELTS

7.1.1.1 IELTS Listening

The IELTS Listening test encompasses a wide range of descriptors, covering all levels from CSE 2 to CSE 8. There is evidence of some progression in difficulty across the test. Section 1 focuses on CSE 2–4. Sections 2 and 3 were judged to be targeted predominantly at CSE 4 with some descriptors at CSE 3. Section 4, consisting essentially of an extended monologue, was judged to cover the widest range of levels from 5 to 8. This is also reflected in the CEFR levels allocated, progressing from A2 in Section 1, through B1 in Sections 2 and 3, and culminating in B2 for Section 4.

The highest number of descriptors was selected from the overall category, but both interaction and exposition are well represented, consistent with the dialogic nature of Sections 1 and 3, and the explanatory nature of the texts throughout the test. It may be noted that there is no obvious progression in difficulty in terms of key information required. All items target specific information within sentences and across sentences in each section with two items targeting information across paragraphs. However, in terms of cognitive processing, the greater difficulty of the later items, particularly in Section 4, can be explained by progression to the more abstract informational content, greater syntactic complexity, less frequent vocabulary and greatly increased length of utterances, meaning that most target information remained within sentences.

Table 10: CSE descriptors allocated to IELTS Listening test by level

Listening sections						
CSE	**1**	**2**	**3**	**4**	**Total**	**%**
1					**0**	0.0
2	4				**4**	14.3
3	3	2	2		**7**	25.0
4	2	6	4		**12**	42.9
5				2	**2**	7.1
6				1	**1**	3.6
7				1	**1**	3.6
8				1	**1**	3.6
9					**0**	0.0
Total	**9**	**8**	**6**	**5**	**28**	**100**

Table 11: CSE descriptors allocated to IELTS Listening test by scale

CSE scales	n	%
Overall	8	28.6
Description	3	10.7
Narration	1	3.6
Exposition	6	21.4
Instruction	2	7.1
Argumentation	1	3.6
Interaction	7	25.0
Total	**28**	**100**

Figure 4: Percentage of CSE descriptors allocated to IELTS Listening test by level

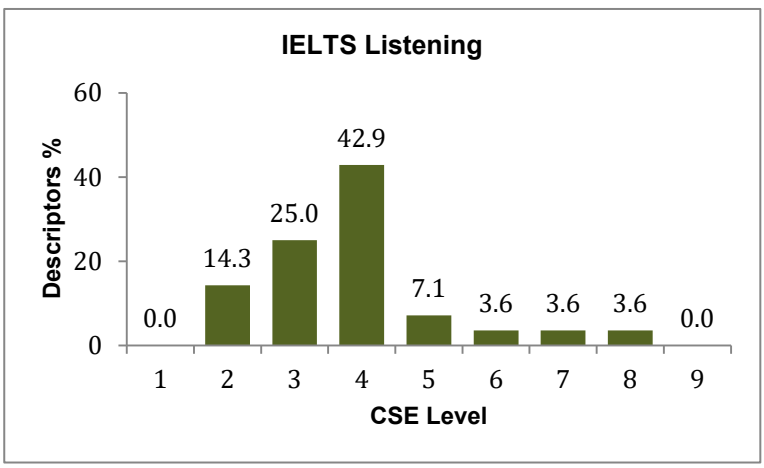


Figure 5: Percentage of CSE descriptors allocated to IELTS Listening test by scale

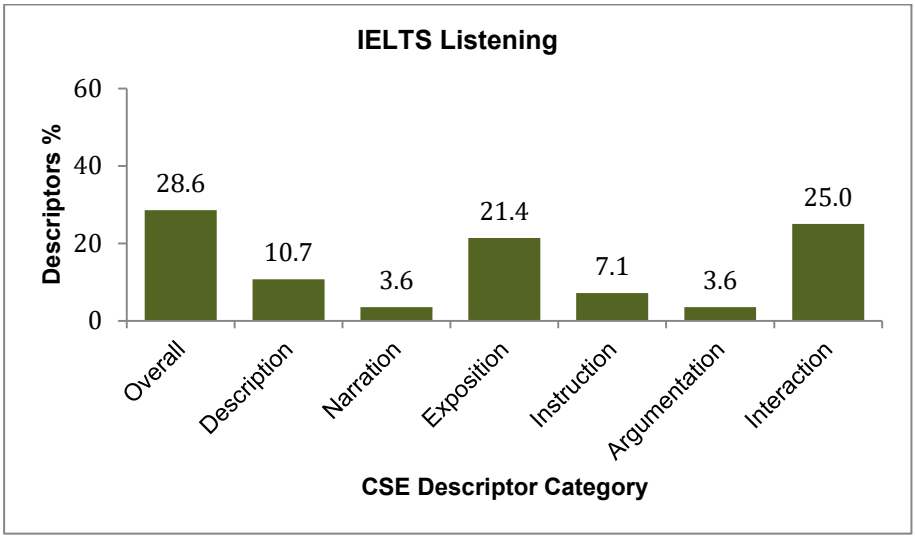
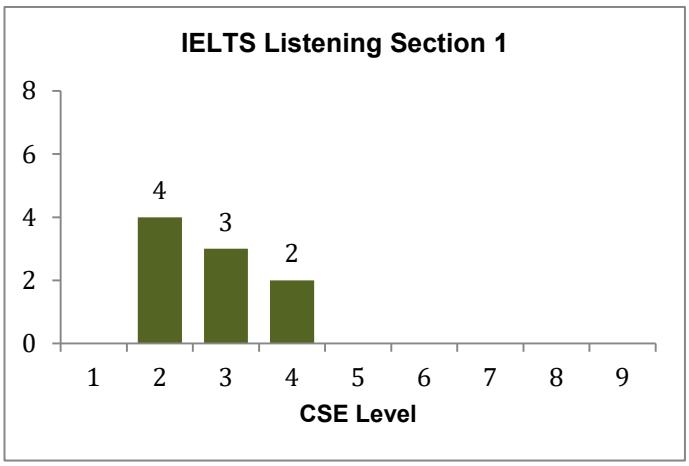
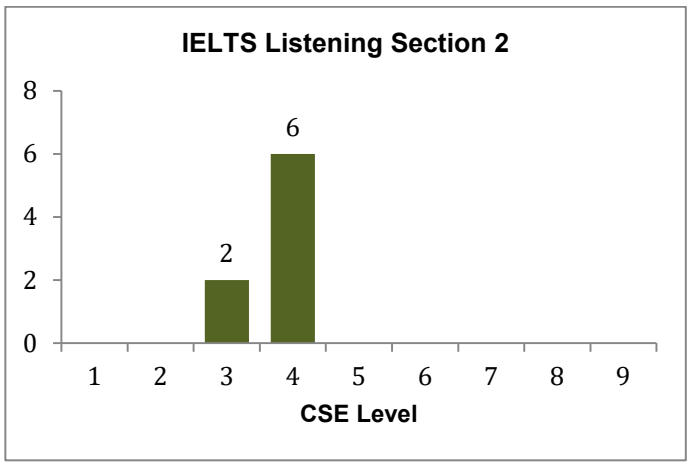
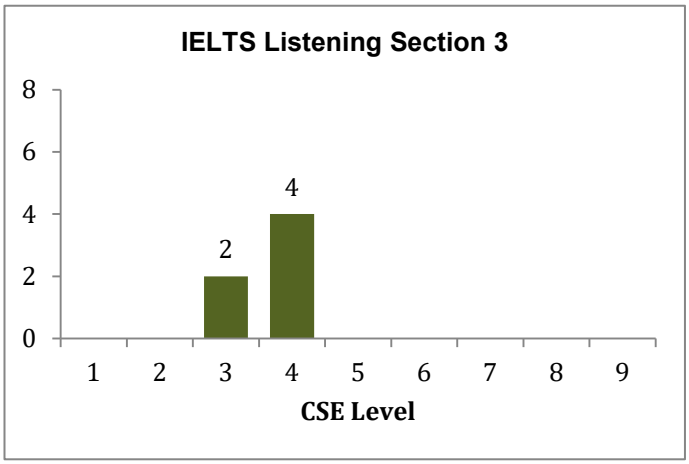
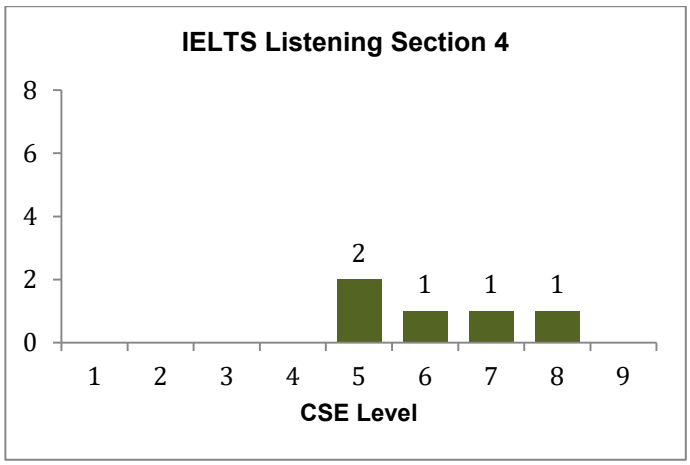


Figure 6: CSE descriptors allocated to IELTS Listening tasks by level









7.1.1.2 IELTS Reading

The IELTS Reading test as a whole covers a range of three CSE levels, with a clear progression across the three sections. Section 1 was considered as matching descriptors only at CSE 6, while Section 2 was matched with descriptors at both CSE levels 6 and 7. Section 3 targets abilities described by CSE 7 and 8. The overall CEFR levels allocated also rose across the test, from B2 in Sections 1 and 2 to C1 in Section 3. In terms of text type, all three texts are both expository and, to some extent, argumentative in nature, and CSE descriptors were chosen only from these two categories, in addition to some descriptors from the overall scale. This is congruent with the fact that the Reading test paper sample is from the academic strand of IELTS. The three sections cover a variety of cognitive process, with each chiefly targeting a different reading style, global expeditious, local careful and global careful reading, respectively. Sections 1 and 2 both target specific information within and across sentences. However, Section 3 additionally targets opinions and main ideas across paragraphs and the text as a whole. Together with the 'mostly abstract' informational content, these are consistent with the higher levels of CSE descriptors selected for this section.

Table 12: CSE descriptors allocated to IELTS Reading test by level

Reading sections					
CSE	**1**	**2**	**3**	**Total**	**%**
1				**0**	0.0
2				**0**	0.0
3				**0**	0.0
4				**0**	0.0
5				**0**	0.0
6	4	2		**6**	42.9
7		3	3	**6**	42.9
8			2	**2**	14.3
9				**0**	0.0
Total	**4**	**5**	**5**	**14**	**100**

Table 13: CSE descriptors allocated to IELTS Reading test by scale

CSE scales	n	%
Overall	3	21.4
Description	0	0.0
Narration	0	0.0
Exposition	7	50.0
Instruction	0	0.0
Argumentation	4	28.6
Interaction	0	0.0
Total	**14**	**100**

Figure 7: Percentage of CSE descriptors allocated to IELTS Reading test by level

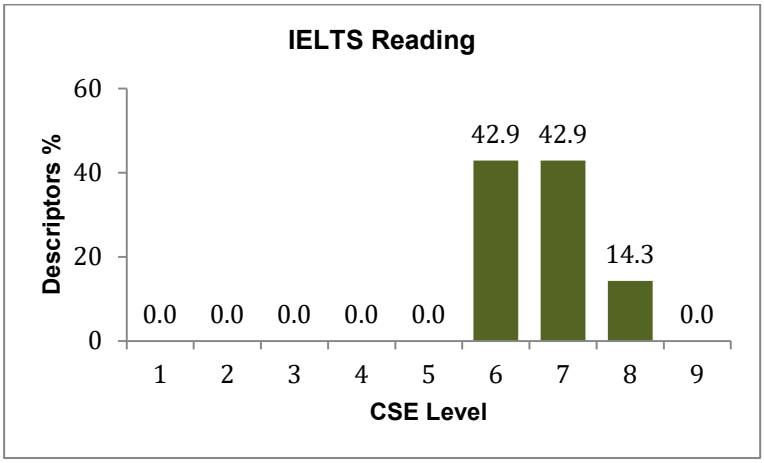


Figure 8: Percentage of CSE descriptors allocated to IELTS Reading test by scale

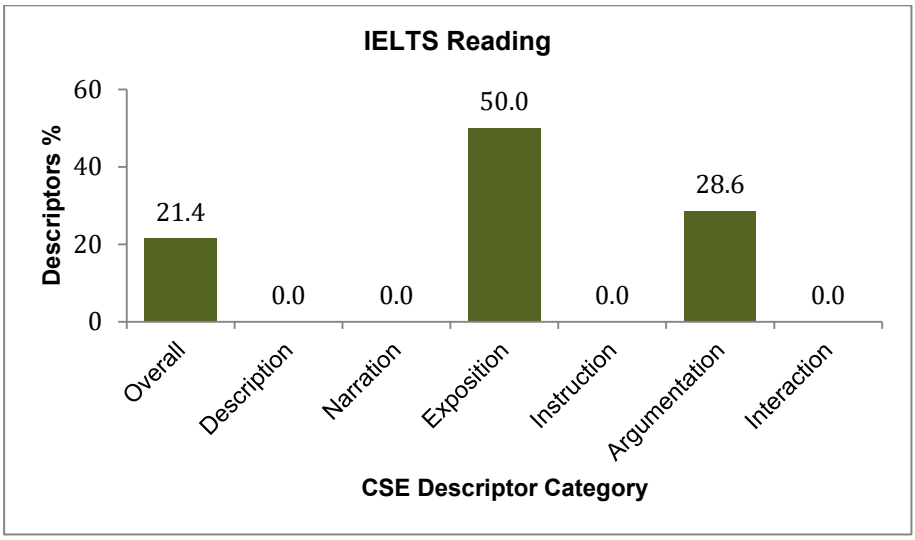
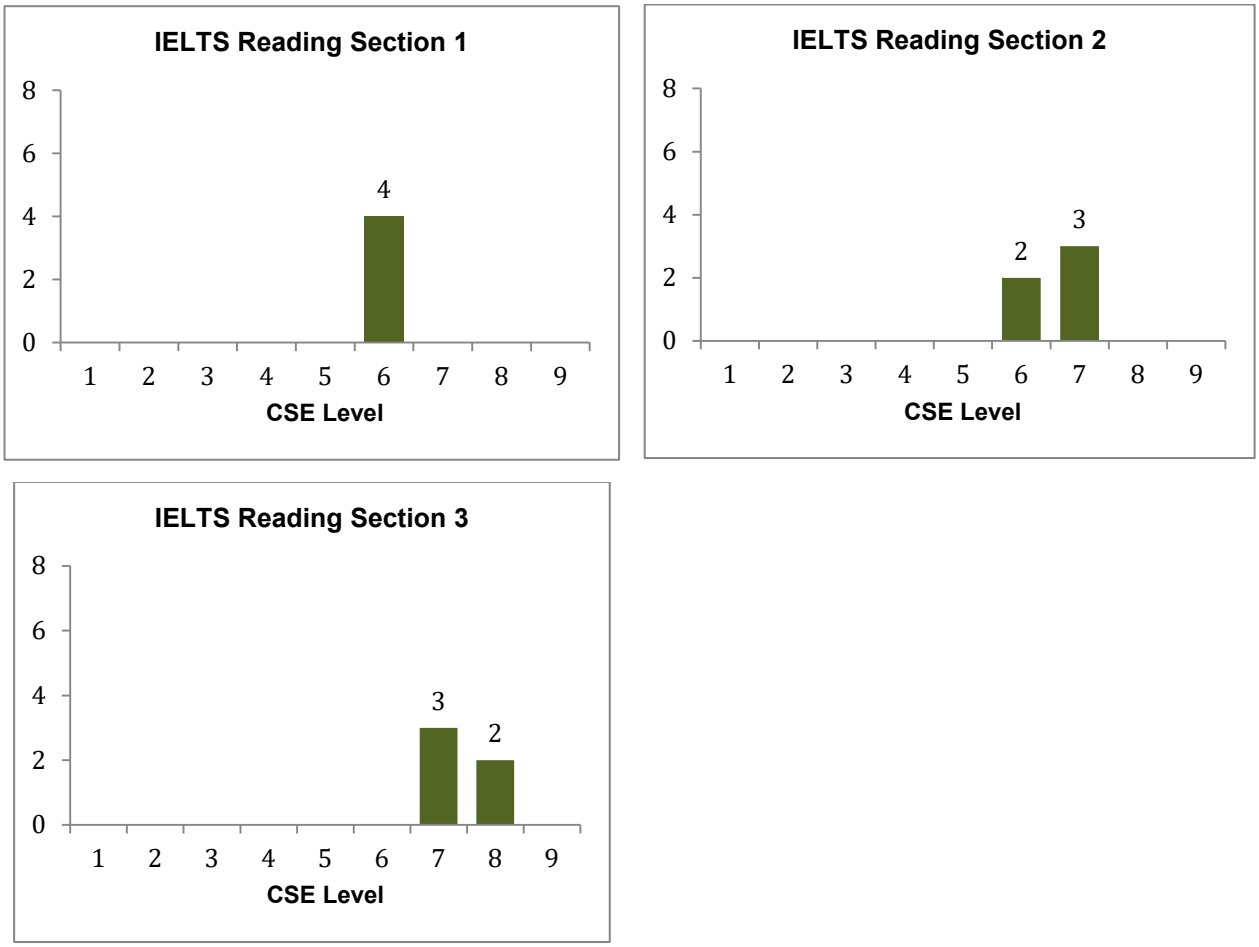


Figure 9: CSE descriptors allocated to IELTS Reading tasks by level



7.1.1.3 IELTS Speaking

The IELTS Speaking test covers a range of five CSE levels, from 3 to 7, with the majority of descriptors selected from CSE levels 5 and 6 (56.7%). The levels of the descriptors rise across the test. While Section 1 is mainly targeted at level 4, Section 2 was matched predominantly with descriptors from level 5, with two descriptors taken from level 4. However, there is a clear distinction between these and Section 3, with descriptors starting at level 6 and including some from level 7. This is also reflected in the jump from CEFR B1 in the first two sections, which require answers of concrete information within the test-taker's personal domain, to B2 in the third section, where more extended output pertaining to more abstract topics in the public domain is required.

Over the test as a whole, a variety of CSE text function categories are dealt with, covering description, exposition, argumentation and interaction, all to a comparable degree. Similarly, all three tasks were judged potentially to cover a broad range of informational functions, dependent on the individual responses of test-takers. However, interactional functions and managing interaction functions were considered to be targeted mainly in Section 3, according to the more abstract and argument-based response necessary. This is also reflected in the CSE descriptors allocated to this task for interaction – one descriptor at level 6 and 2 at level 7.

Table 14: CSE descriptors allocated to IELTS Speaking test by level

Speaking sections					
CSE	**1**	**2**	**3**	**Total**	**%**
1				**0**	0.0
2				**0**	0.0
3	1			**1**	3.8
4	4	2		**6**	23.1
5		7		**7**	26.9
6			8	**8**	30.8
7			4	**4**	15.4
8				**0**	0.0
9				**0**	0.0
Total	**5**	**9**	**12**	**26**	**100**

Table 15: CSE descriptors allocated to IELTS Speaking test by scale

CSE scales	n	%
Overall	6	23.1
Description	4	15.4
Narration	0	0.0
Exposition	5	19.2
Instruction	0	0.0
Argumentation	6	23.1
Interaction	5	19.2
Total	**26**	**100**

Figure 10: Percentage of CSE descriptors allocated to IELTS Speaking test by level

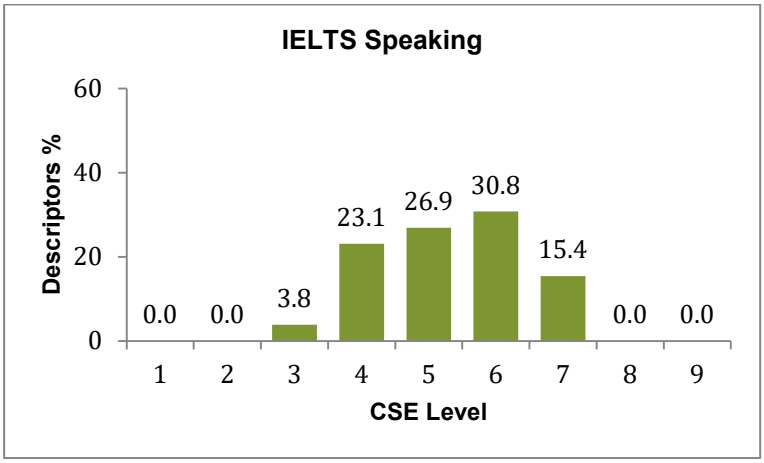


Figure 11: Percentage of CSE descriptors allocated to IELTS Speaking test by scale

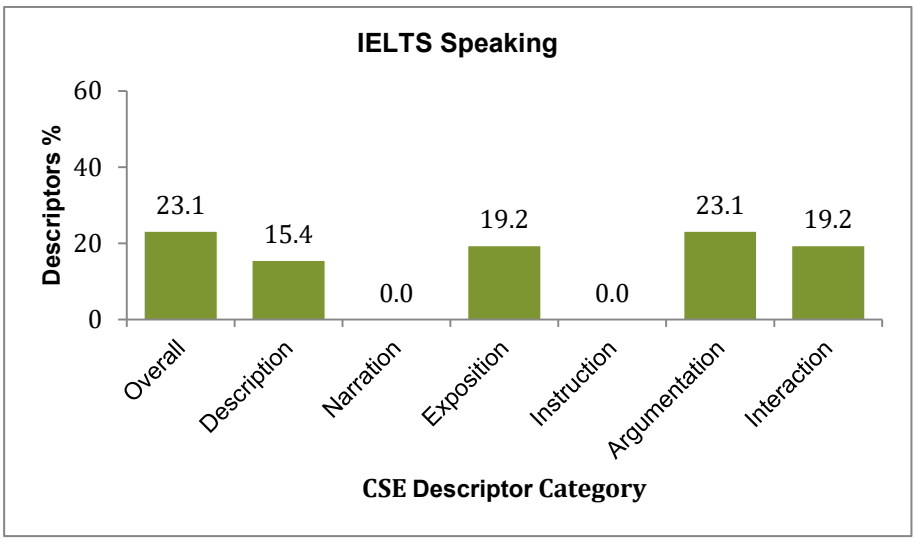
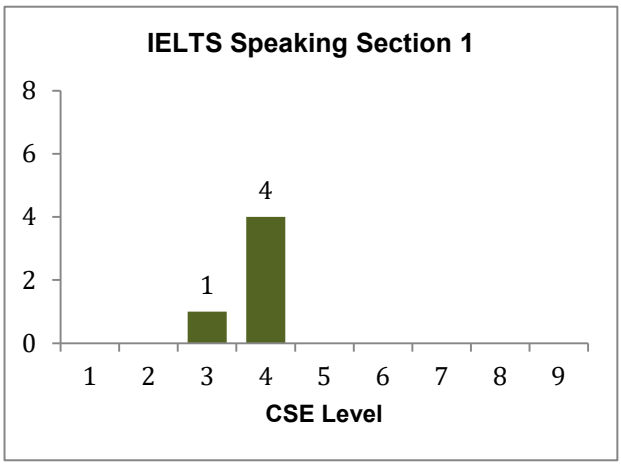
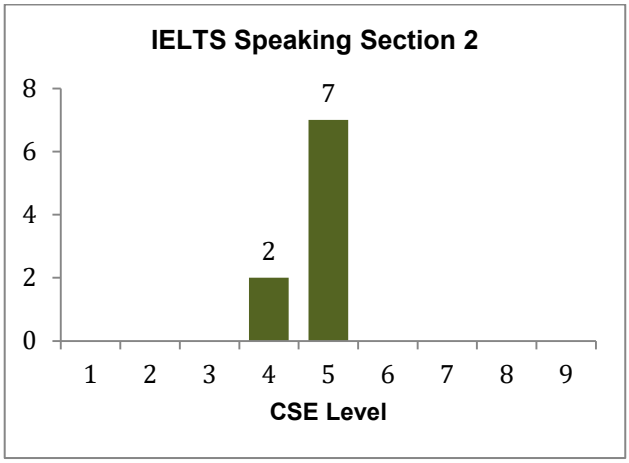
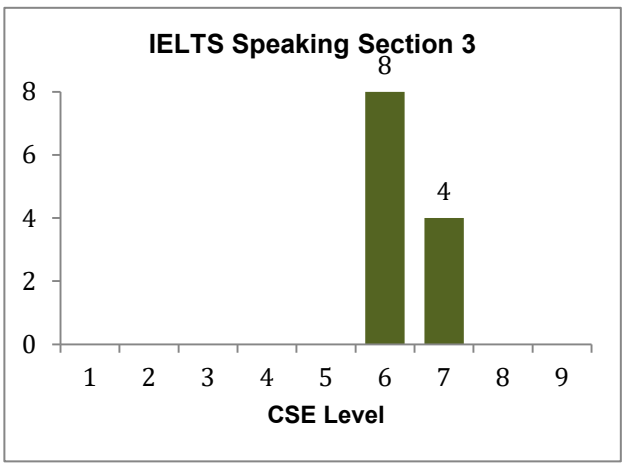


Figure 12: CSE descriptors allocated to IELTS Speaking tasks by level







7.1.1.4 IELTS Writing

The IELTS Academic Writing paper as a whole covers a range of four CSE levels, with CSE 5 the level from which descriptors were more commonly chosen. Nevertheless, it should be noted here that the writing component comprises only two sections. Therefore, relatively few CSE descriptors (a count of 12 in total) could be applied to this component in comparison with other sections of IELTS, although all categories of descriptor are represented except for interaction. Descriptors for Section 1, the description of a process using graphical input, were shared equally between CSE 3 and 5, but there were no descriptors for CSE 4. While such a distribution of levels may seem unusual, it can be attributed to the specificity of the task itself in the occupational domain, and the specific contextual parameters included as features of the CSE descriptors, for example, references to film and familiar places. Section 2, an opinion-based essay in the educational domain, could be matched with a larger number of descriptors (8) across three levels (CSE 4–6). The argumentative discourse mode identified was also reflected by the selection of three descriptors from the argumentation category and two from exposition. The CEFR levels allocated for the two sections are B1 and B2, respectively, in line with the difference in ability specified by the CSE levels. Similarly, the demands of the Section 2 task differ markedly in terms of greater length of response, wider range of functions, and more abstract nature of information.

Table 16: CSE descriptors allocated to IELTS Writing test by level

Writing sections				
CSE	**1**	**2**	**Total**	**%**
1			**0**	0.0
2			**0**	0.0
3	2		**2**	16.7
4		2	**2**	16.7
5	2	3	**5**	41.7
6		3	**3**	25.0
7			**0**	0.0
8			**0**	0.0
9			**0**	0.0
Total	**4**	**8**	**12**	**100**

Table 17: CSE descriptors allocated to IELTS Writing test by scale

CSE scales	n	%
Overall	2	16.7
Description	1	8.3
Narration	1	8.3
Exposition	4	33.3
Instruction	1	8.3
Argumentation	3	25.0
Interaction	0	0.0
Total	**12**	**100**

Figure 13: Percentage of CSE descriptors allocated to IELTS Writing test by level

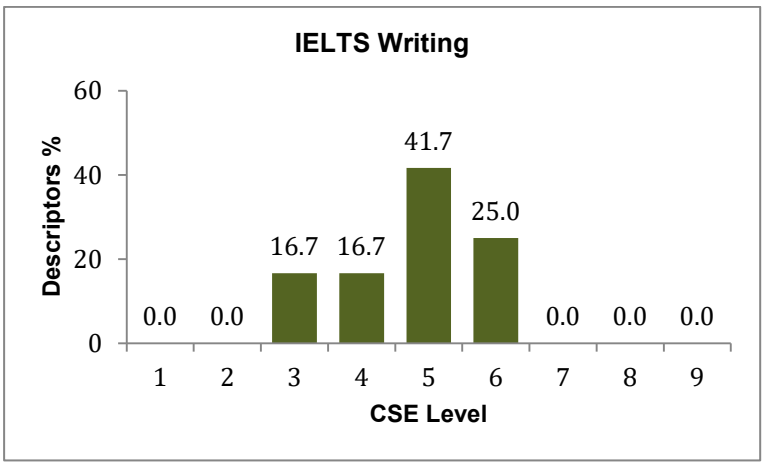


Figure 14: Percentage of CSE descriptors allocated to IELTS Writing test by scale

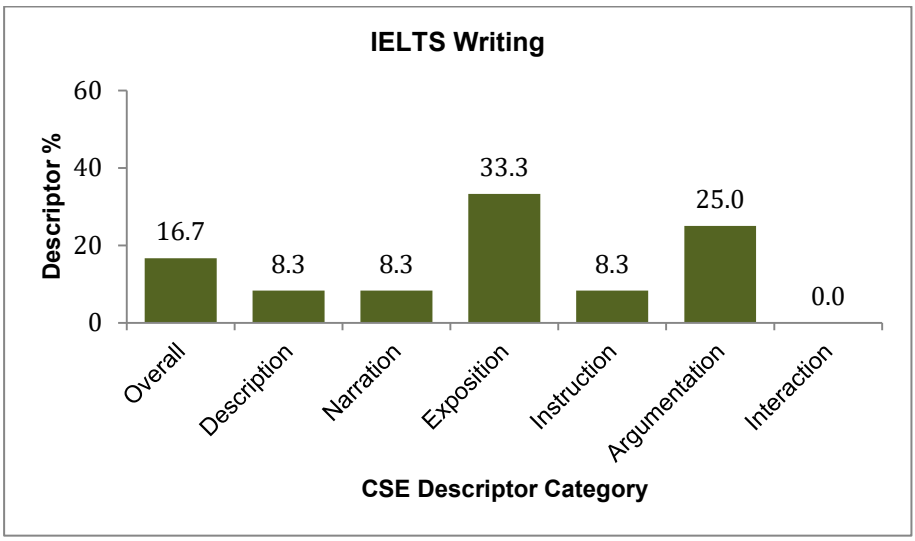
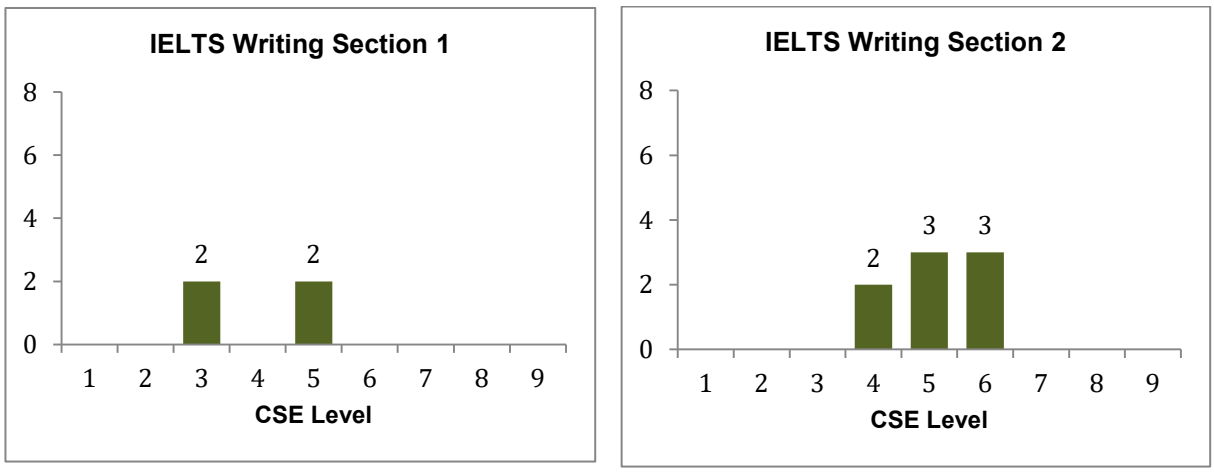


Figure 15: CSE descriptors allocated to IELTS Writing tasks by level



7.1.2 Aptis

7.1.2.1 Aptis Listening

For convenience and ease of analysis, the Aptis listening items are divided here according to the intended CEFR level of the items. As mentioned above, explicit CEFR specifications are inherent in the design of the Aptis test. Thus, the four task groupings shown below, 1–5, 6–10, 11–14 and 15–17, are designed to correspond to CEFR levels A1, A2, B1 and B2, respectively. With the exception of one item (Task 2 Item 1), the intended levels matched those of the expert judges.

The Aptis Listening test spans a wide range of different CSE levels, from 2 to 7, although there are no descriptors for level 6. Tasks 1–5 are targeted exclusively at CSE level 2, and involve picking out key words in the listening text, meaning that cognitive processing is lexical search at a careful local level, and dealing with concrete information on daily topics. Tasks 6–10 were matched with descriptors split between CSE 2 and CSE 3. As can be seen in Appendix C, while sharing many of the cognitive processing characteristics of the previous tasks, the nature of information is less concrete and a greater variety of topics (travel, leisure) are dealt with.

Tasks 11–14 target only CSE 4. The level of cognitive processing here was identified as increasing to meaning construction, where information is integrated across utterances to build a mental model of the text. Accordingly, the processing targeted in these tasks is a combination of careful local where specific information is required (Tasks 11–13) and expeditious global in comprehending the gist of the text (Task 14). Tasks 15–17 were matched with descriptors from CSE 3, CSE 5 and CSE 7. This distribution can be attributed to the perceived difference in processing for Task 17 and the other tasks in this section. While the key information in Task 17 is across sentences, the key information for Tasks 15 and 16 is at text level, with a higher level of cultural specificity and content knowledge required.

The majority (56.8%) of descriptors used were selected from the overall category, with exposition the next largest category, followed by interaction, reflecting the balance of monologue and dialogue texts and tasks.

Table 18: CSE descriptors allocated to Aptis Listening test by level

CSE	Tasks 1–5	Tasks 6–10	Tasks 11–14	Tasks 15–17	Total	%
1					**0**	0.0
2	9	6			**15**	40.5
3		8		1	**9**	24.3
4			7		**7**	18.9
5				2	**2**	5.4
6					**0**	0.0
7				4	**4**	10.8
8					**0**	0.0
9					**0**	0.0
Total	**9**	**14**	**7**	**7**	**37**	**100**

Table 19: CSE descriptors allocated to Aptis Listening test by scale

CSE scales	n	%
Overall	21	56.8
Description	1	2.7
Narration	1	2.7
Exposition	8	21.6
Instruction	1	2.7
Argumentation	2	5.4
Interaction	3	8.1
Total	**37**	**100**

Figure 16: Percentage of CSE descriptors allocated to Aptis Listening test by level

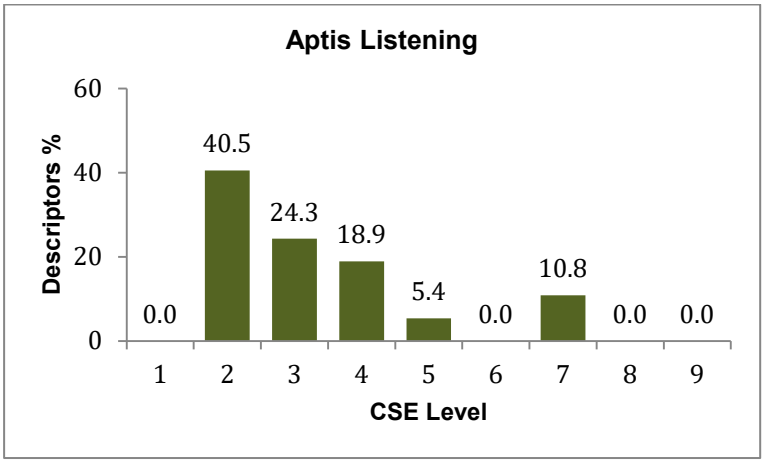


Figure 17: Percentage of CSE descriptors allocated to Aptis Listening test by scale

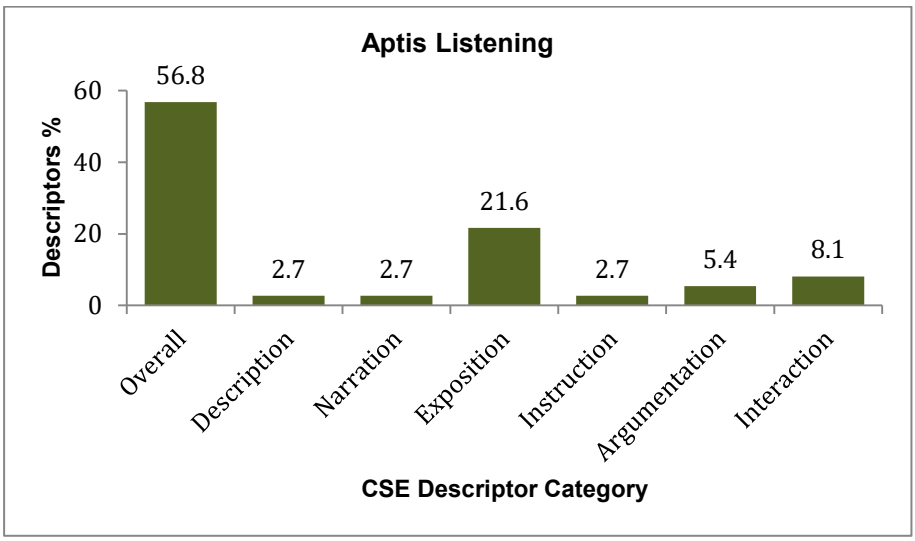
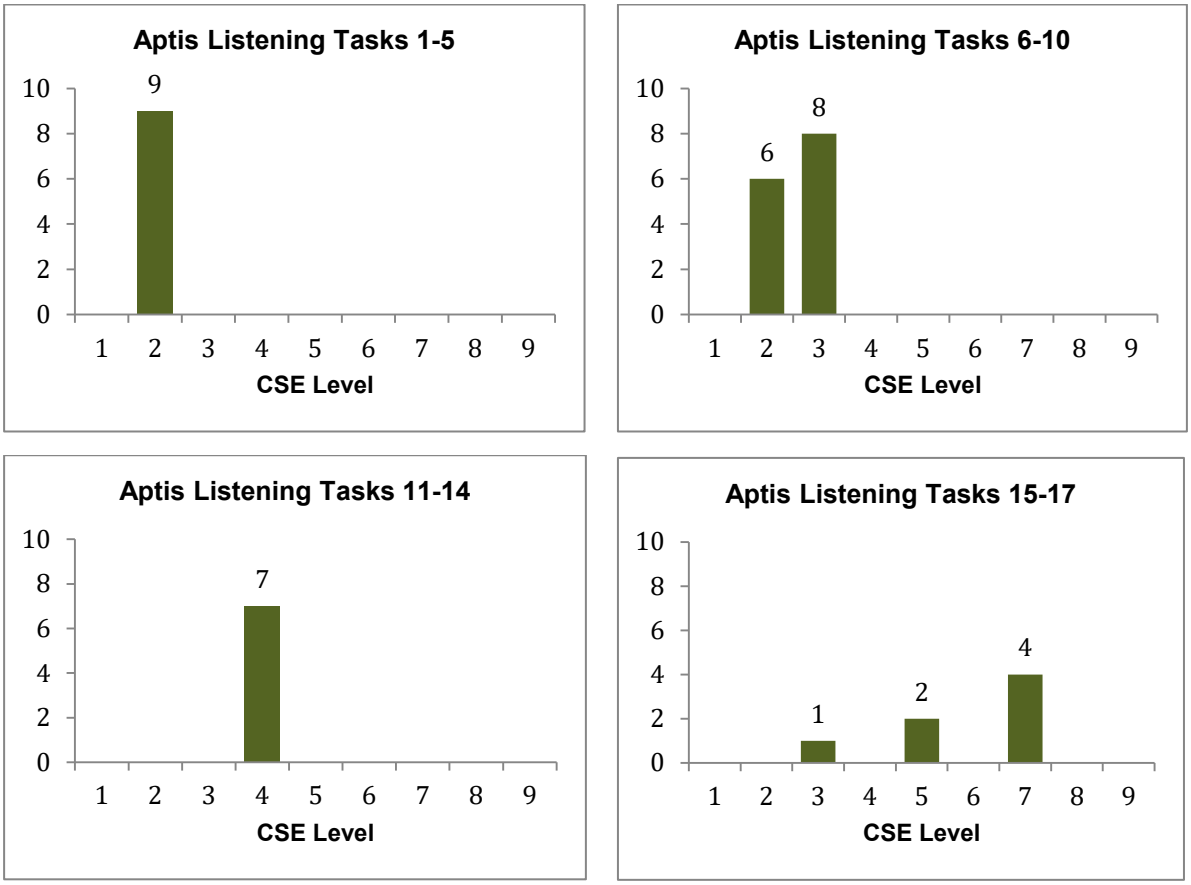


Figure 18: CSE descriptors allocated to Aptis Listening tasks by level



7.1.2.2 Aptis Reading

The Aptis Reading test covers a range of five CSE bands, from CSE 2 to CSE 6. The progression across tasks is very clear, with each task targeted at a single distinct level. Task 1 was identified as CSE 2. Tasks 2 and 3, which comprise a single section on the Aptis test, was matched with CSE 3. Task 4 was judged to be at CSE 4, while Task 5 was matched with both CSE level 5 and level 6. This progression in the CSE is consistent with the progression in CEFR levels across these tasks, with Task 1 at A1, Tasks 2 and 3 at A2, Task 4 at B1 and Task 5 at B2. These CEFR levels as allocated by the expert judges, are also a complete match with the targeted CEFR levels within the Aptis Reading test specifications described in Section 6.2 above. In terms of the different CSE scales identified, the overall scale is the largest category, but narration, exposition and argumentation are also represented with more than one descriptor in each. In terms of cognitive processing and task features, it can be seen that each task has distinct characteristics, accounting for the clear differences in descriptor levels attributed to them. Task 1 targets specific information within sentences, employing careful global reading at the level of lexical access. Tasks 2 and 3 target text structure across sentences, utilising careful global reading for text level representation. Task 4 targets writer's attitude across sentences, using careful global reading to establish propositional meaning. Task 5 involves a longer text, targeting text structure across paragraphs, engaging expeditious global reading to build a mental model of the text.

Table 20: CSE descriptors allocated to Aptis Reading test by level

Reading tasks							
CSE	**1**	**2**	**3**	**4**	**5**	**Total**	**%**
1						**0**	0.0
2	4					**4**	26.7
3		2	2			**4**	26.7
4				3		**3**	20.0
5					2	**2**	13.3
6					2	**2**	13.3
7						**0**	0.0
8						**0**	0.0
9						**0**	0.0
Total	**4**	**2**	**2**	**3**	**4**	**15**	**100**

Table 21: CSE descriptors allocated to Aptis Reading test by scale

CSE scales	n	%
Overall	5	33.3
Description	0	0.0
Narration	3	20.0
Exposition	3	20.0
Instruction	0	0.0
Argumentation	3	20.0
Interaction	1	6.7
Total	**15**	**100**

Figure 19: Percentage of CSE descriptors allocated to Aptis Reading test by level

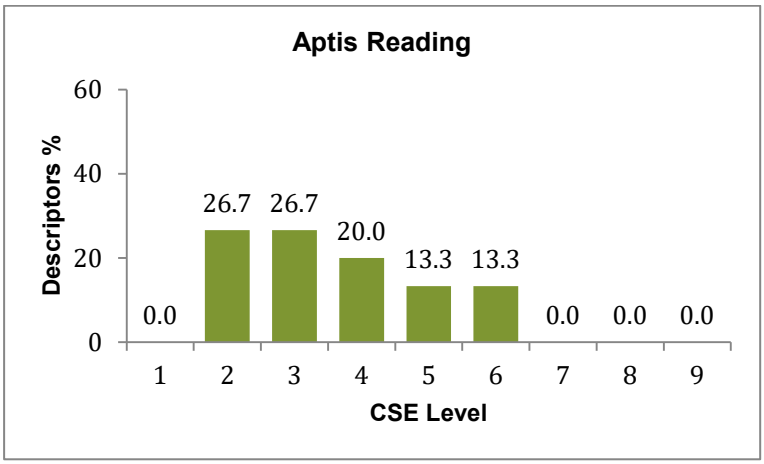


Figure 20: Percentage of CSE descriptors allocated to Aptis Reading test by scale

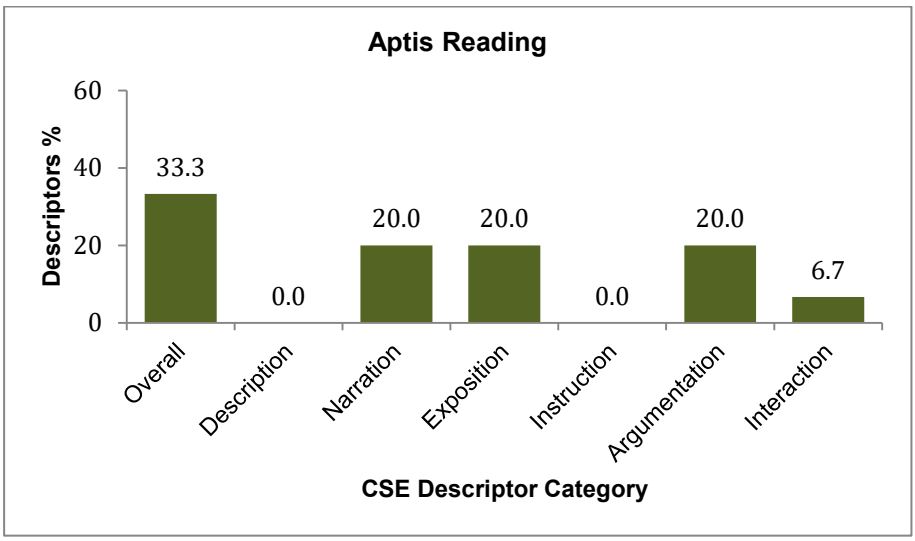
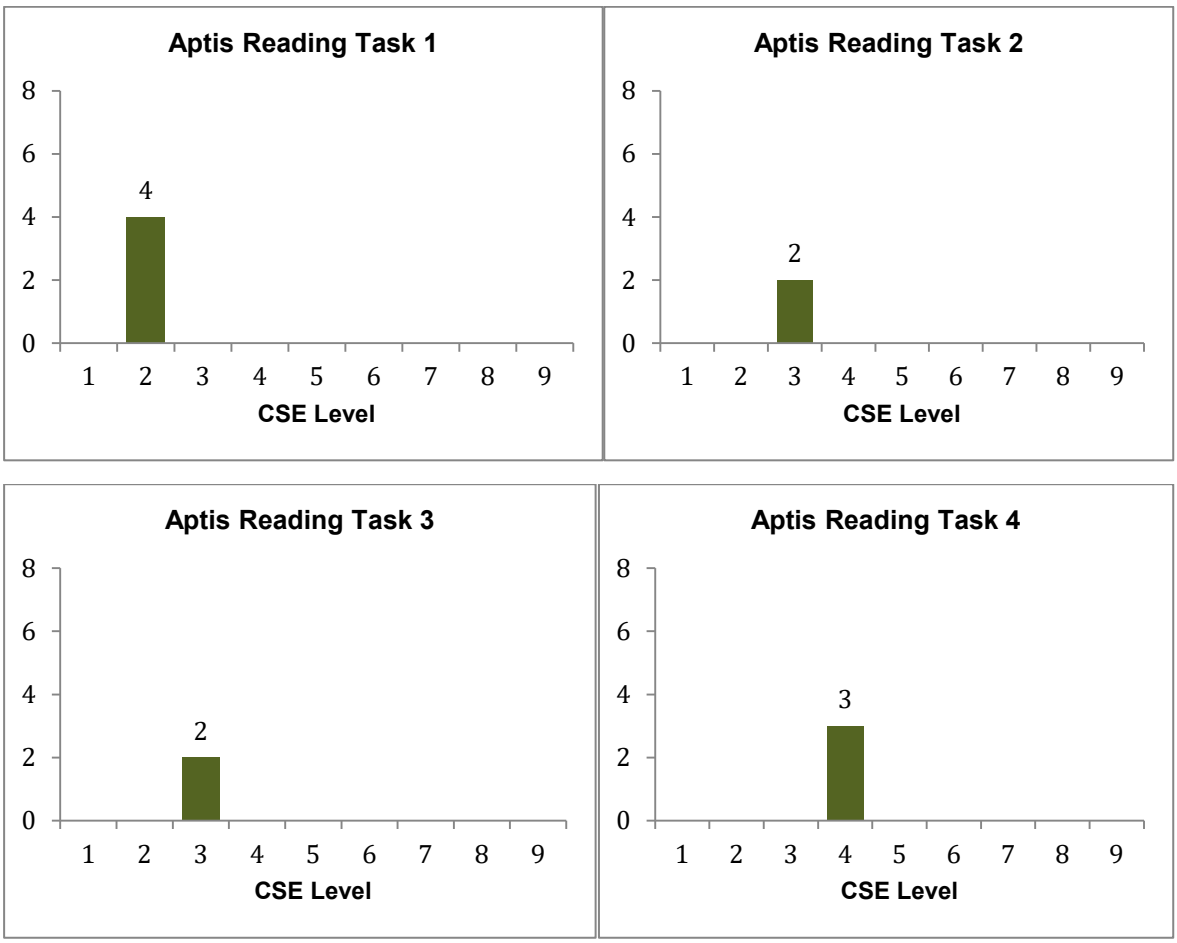
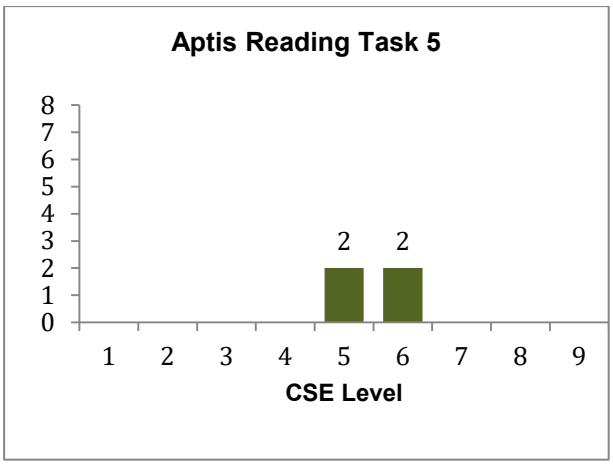


Figure 21: CSE descriptors allocated to Aptis Reading tasks by level





7.1.2.3 Aptis Speaking

The Aptis Speaking test covers a range of four levels, from CSE 3 to CSE 6. In terms of the levels of CSE descriptors identified by the expert judges, there is evident progression in difficulty across the four parts, but also a degree of overlap between all adjacent parts. While Part 1 was identified as targeting CSE 3 and Part 2 CSE 3 and 4, Part 3 was felt to target CSE 4 and 5. Part 4, the extended speaking turn, was matched with three CSE levels, CSE 4–6. The CEFR bands allocated were A2 for Part 1 and B1 for the remaining parts. This is in line with the Aptis test design for Tasks 1–3, but not for Task 4, which is designed to be a B2 task. The Speaking test as a whole was matched with descriptors from across the CSE scales, except for instruction and interaction. The lack of a match with interaction descriptors may well be a feature of the delivery of Aptis, in which the test-taker interacts with the computer rather than a human interlocutor. In terms of the features of the task, no interactional or managing interaction functions were identified across the whole test, and the personal domain was identified for all items except Task 4 Item 3. However, there were differences in the number of informational functions among tasks, increasing in number across Tasks 1–3. Despite the increased length of the turn required in Task 4, there were fewer perceived differences between Task 4 and Task 3, with a smaller number of informational functions for Task 4, even though it was matched with higher level CSE descriptors by the expert judges.

Table 22: CSE descriptors allocated to Aptis Speaking test by level

Speaking tasks						
CSE	**1**	**2**	**3**	**4**	**Total**	**%**
1					**0**	0.0
2					**0**	0.0
3	2	4			**6**	28.6
4		2	3	1	**6**	28.6
5			2	4	**6**	28.6
6				3	**3**	14.3
7					**0**	0.0
8					**0**	0.0
9					**0**	0.0
Total	**2**	**6**	**5**	**8**	**21**	**100**

Table 23: CSE descriptors allocated to Aptis Speaking test by scale

CSE scales	n	%
Overall	8	38.1
Description	4	19.0
Narration	2	9.5
Exposition	2	9.5
Instruction	0	0.0
Argumentation	5	23.8
Interaction	0	0.0
Total	**21**	**100**

Figure 22: Percentage of CSE descriptors allocated to Aptis Speaking test by level

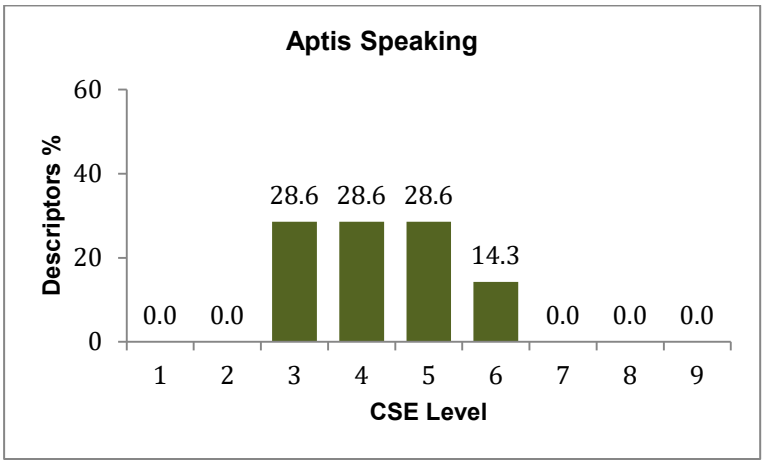


Figure 23: Percentage of CSE descriptors allocated to Aptis Speaking test by scale

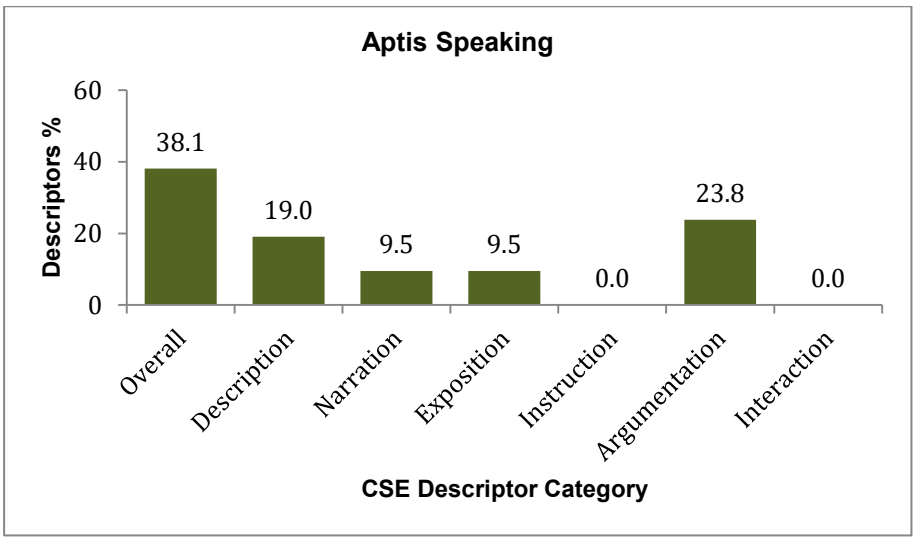
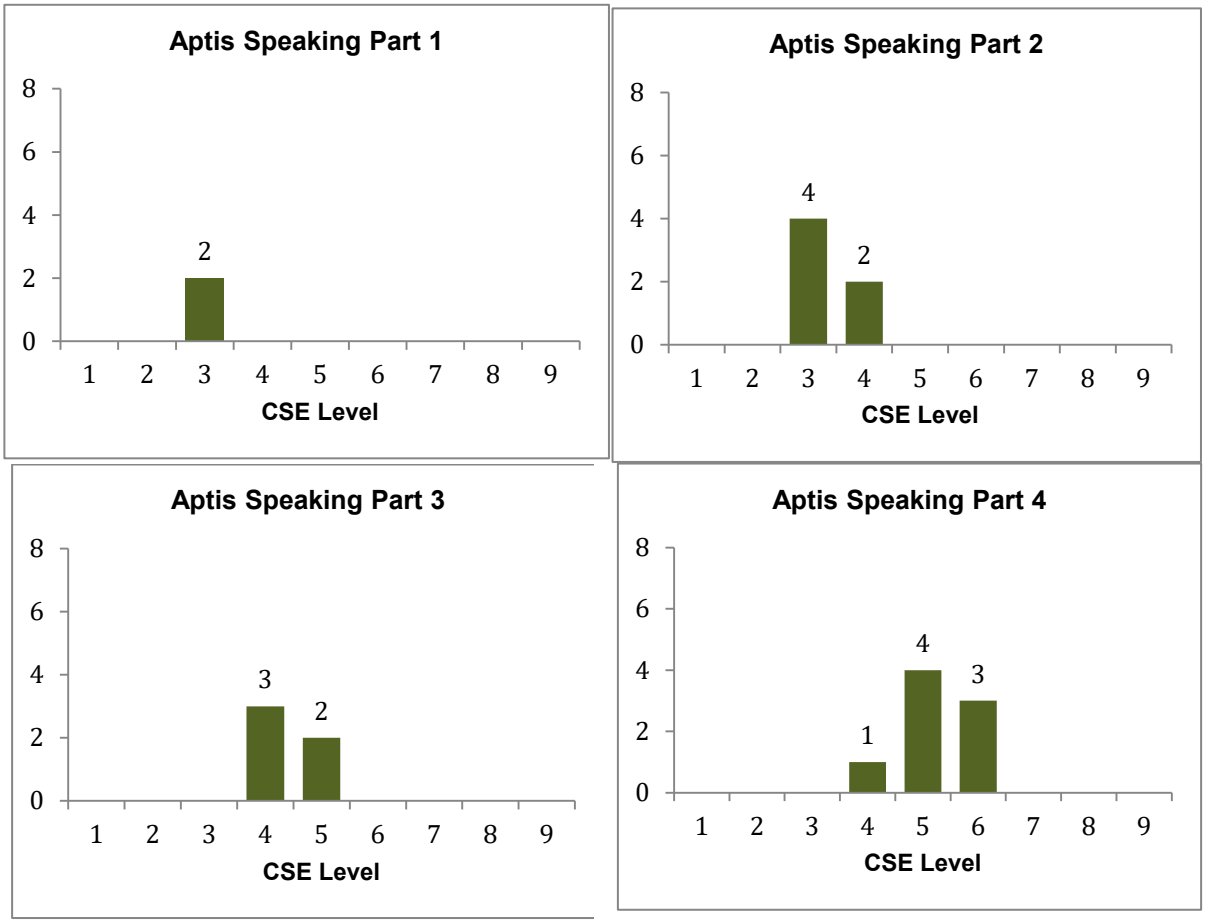


Figure 24: CSE descriptors allocated to Aptis Speaking tasks by level



7.1.2.4 Aptis Writing

The Aptis Writing test covers a range of six CSE levels, from CSE 1 through to CSE 6, with the majority of descriptors at levels 3 and 4. For Part 1, in which only single words or phrases are required from test candidates, only CSE 1 was found to be appropriate. Part 2, which requires a single short response in full sentences, was associated predominantly with CSE 3, while Part 3, which requires multiple short responses, was identified as being at CSE 3 and 4. Part 4, which requires candidates to write a short informal text followed by a longer, more formal text, was identified as CSE 4–6, together with just one descriptor at CSE 3. This progression across tasks was also reflected in the CEFR levels allocated, from A1 for Part 1 to B2 to Part 4, matching the levels detailed in the Aptis test specifications. Apart from the overall scale, the category of descriptors common to all parts was exposition, with argumentation emerging as a significant category at higher levels in Part 3 and Part 4. Across the whole test, certain features were deemed to be constant, such as the personal domain and mostly concrete nature of information, together with the generality and neutrality of the content, potentially contributing to the inclusion of lower CSE levels for the later tasks. However, differences are apparent and may account for the gradual increase in task level, particularly the increasing variety of functions, from basic functions of providing personal and routine information in Part 1, to much greater complexity in Task 4, such as developing an argument.

Table 24: CSE descriptors allocated to Aptis Writing test by level

Writing tasks						
CSE	**1**	**2**	**3**	**4**	**Total**	**%**
1	4				**4**	13.3
2		1			**1**	3.3
3		3	5	1	**9**	30.0
4			5	6	**11**	36.7
5				3	**3**	10.0
6				2	**2**	6.7
7					**0**	0.0
8					**0**	0.0
9					**0**	0.0
Total	**4**	**4**	**10**	**12**	**30**	**100**

Table 25: CSE descriptors allocated to Aptis Writing test by scale

CSE scales	n	%
Overall	10	33.3
Description	1	3.3
Narration	3	10.0
Exposition	8	26.7
Instruction	0	0.0
Argumentation	6	20.0
Interaction	2	6.7
Total	**30**	**100**

Figure 25: Percentage of CSE descriptors allocated to Aptis Writing test by level

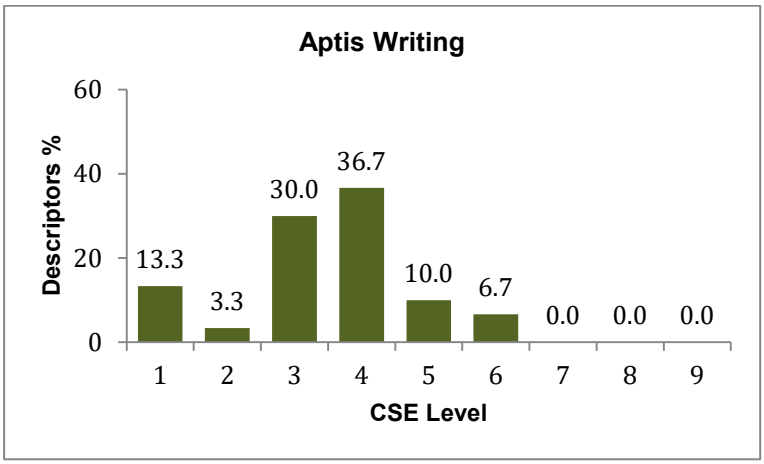


Figure 26: Percentage of CSE descriptors allocated to Aptis Writing test by scale

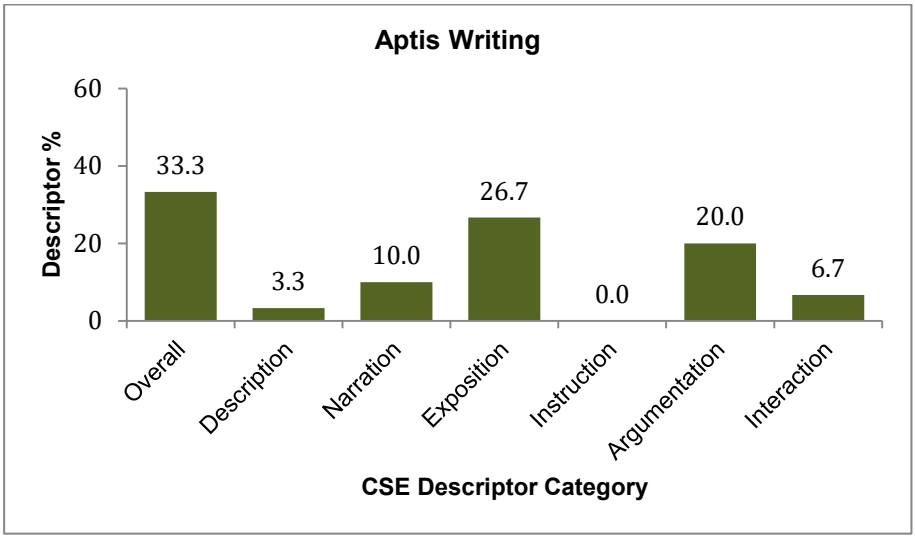
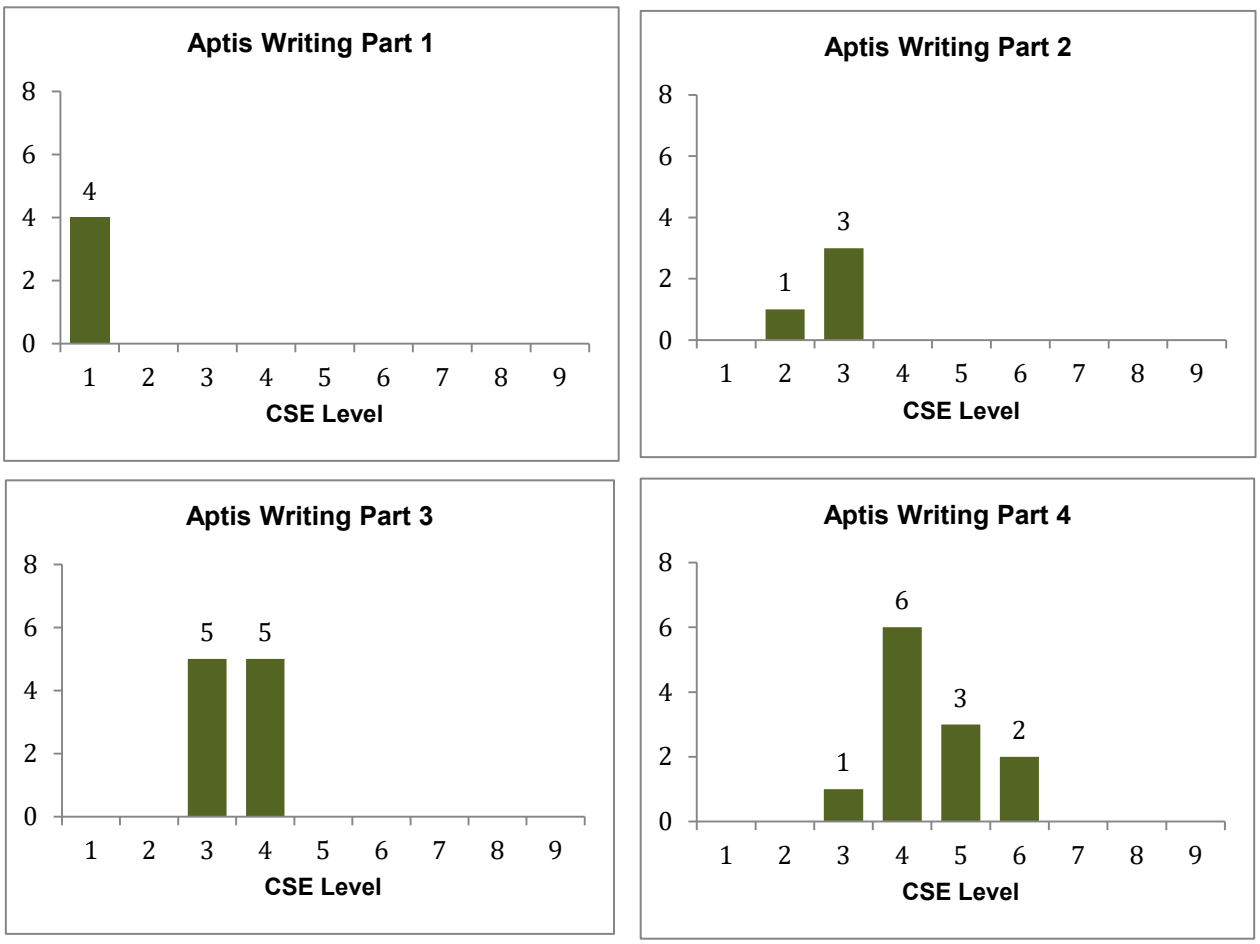


Figure 27: CSE descriptors allocated to Aptis Writing tasks by level



7.2 Standard-setting panels

7.2.1 Overview

Three separate panels were organised, focusing on different skills. Listening was held first and treated as a pilot to trial the methodology. Reading was carried out several months later as the second panel, as both of these receptive skills components utilised the same pair of test-centred standard setting methods, as described in Section 5. Finally, Speaking and Writing were combined and run with the same panellists over a five-day period. Appendix B contains the schedules for each panel. Certain aspects of the panel procedures were shared and will be described in brief here. The following sections will then present the results for each skill component separately.

All panels followed a similar approach in having self-study preparation booklets to familiarise participants with the CSE. The face-to-face panels then followed a similar pattern, beginning with one day of training to ensure participants had a shared interpretation of the CSE levels. These training sessions were led by Working Group members experienced in standard setting. Each panel carried out standard setting for Aptis first, and then proceeded to repeat the process with IELTS. The initial qualitative review by the Working Group of the constructs targeted by the tests had led to a decision to focus on CSE levels 3–6 for Aptis and 4–7 for IELTS, with the aim to set cutscores for each of the levels in those ranges. However, during the initial panel for Listening, it quickly became clear that it would be possible to cover levels 3–7 for Aptis and 4–8 for IELTS. This range of target levels was then fixed and used for all subsequent panels.

7.2.2 Listening

7.2.2.1 Introduction

As described in Section 5.3.4, the Listening panel was carried out as the first of a series of standard-setting panels planned to cover all of the four skill components of the Aptis and IELTs exams, which were the subject of this linking project. The procedures employed followed the outline given in the Methodology section and which is also reflected in the schedule of activities presented in Appendix B. A total of 16 panellists took part in the Listening standard-setting panel (for an overview of the panellists and also their feedback on the whole standard setting process, see Section 7.2.6 on procedural validity).

As already noted, the Basket Method was applied first as a way of helping participants to conceptualise the difficulty of the items in terms of an appropriate CSE level. After making their judgements using the Basket Method, participants were presented with normative feedback so that they were aware of the position of their judgements in relation to the other panellists and they engaged in discussion on the rationale for their judgements. Following the Basket Method, two rounds of judgements were carried with the Modified Angoff Method. Participants estimated the number of test-takers who would correctly answer each item out of a group of 100 minimally competent candidates at the particular CSE level under consideration at the time. Normative feedback on judgements and empirical item feedback from a live administration of the test under consideration were presented after round one. After discussion on the rationale for their decisions, panellists then carried out a second full round of judgements on all items in the test. The following sections summarise the results, focusing on the analysis of the second round of Angoff judgements, as these judgements were intended from the outset as the final criteria for setting cutoff estimates from the standard-setting panel.

Results were analysed using both Classical Testing Theory (CTT) and multi-facet Rasch model (MFRM) approaches, as described further below.

7.2.2.2 Aptis Listening

7.2.2.2.1 CTT results

As noted, the Basket Method was not intended for setting cutoffs. However, it is useful to report the average level estimates for items from panellists across the test. Table 26 presents an overview of the mean CSE level judgements for each item in the Aptis Listening test by all 16 panellists. The average item level for Aptis Listening is thus 4.75, and items ranged in level estimates from a high of 6.81 to a low of 2.81.

Table 26: Overview of Basket Method judgements for Aptis Listening Test

Mean	SD	Max	Min
4.75	1.53	6.81	2.83

Table 27 presents the cutoff estimates for round two Modified Angoff judgements for each CSE level targeted for Aptis Listening. To estimate cutoffs from Modified Angoff probability judgements, the estimates for each judge are first averaged, deriving a mean percentage-correct estimate across all items for each judge. The mean for each judge is the cutoff estimate for the targeted CSE level for that judge. For example, in Table 27, the mean percentage correct estimate for CSE 3 for R1 (rater 1, the first judge) is 28.8%. This is interpreted as the minimum score a test-taker would need to achieve to demonstrate a proficiency level of CSE 3, and to have crossed the threshold from level CSE 2. For Aptis, a score of 28.8 is then translated into a scale score of 14.4 scale score points on the 0–50 reported score scale used by Aptis (see O'Sullivan & Dunlea, 2015 for details of the scoring and reporting system). The cutoffs for CSE 4, CSE 5, CSE 6, and CSE 7 for Rater 1 are displayed in the subsequent columns. The cutoff estimate for the test from the round two panel judgements is the mean of the cutoff estimates for all judges at that level. The cutoff for CSE 3 from round two panel judgements for the Aptis Listening test is thus 28.88%, or 14.44 points, on the reporting scale of 0–50. Two more statistics are shown at the bottom of each column in Table 27, the standard deviation of the cutoff estimates and the standard error of the cutscore (SE_c) are also shown at the bottom of the table. These two statistics will be discussed further under the section on internal validity of the cutscore estimates.

Table 27: Aptis Listening Angoff Round 2 judgements

	CSE 3		CSE 4		CSE 5		CSE 6		CSE 7	
Rater	**%**	**Scale score**	**%**	**Scale score**	**%**	**Scale score**	**%**	**Scale score**	**%**	**Scale score**
R1	28.8	14.4	43.2	21.6	57.2	28.6	72	36	87.6	43.8
R2	36	18	48.4	24.2	60.8	30.4	71.2	35.6	82	41
R3	28	14	39.6	19.8	56.4	28.2	72.4	36.2	88.4	44.2
R4	28.4	14.2	40.8	20.4	53.2	26.6	64	32	74.8	37.4
R5	43.6	21.8	53.6	26.8	63.6	31.8	73.2	36.6	80.8	40.4
R6	23.6	11.8	44	22	58.8	29.4	76.8	38.4	89.2	44.6
R7	20	10	35.6	17.8	54.8	27.4	74	37	87.2	43.6
R8	23.2	11.6	40	20	66.4	33.2	81.2	40.6	89.2	44.6
R9	28	14	43.2	21.6	65.6	32.8	83.2	41.6	92.8	46.4
R10	26	13	46.4	23.2	64.8	32.4	78	39	88.8	44.4
R11	28.4	14.2	42.4	21.2	57.2	28.6	70	35	79.2	39.6
R12	20.4	10.2	36	18	51.6	25.8	72	36	83.2	41.6

	CSE 3		CSE 4		CSE 5		CSE 6		CSE 7	
Rater	**%**	**Scale score**	**%**	**Scale score**	**%**	**Scale score**	**%**	**Scale score**	**%**	**Scale score**
R13	27.2	13.6	42.4	21.2	66.4	33.2	80.4	40.2	90.8	45.4
R14	28.8	14.4	43.6	21.8	55.6	27.8	68.8	34.4	81.2	40.6
R15	35.2	17.6	42.4	21.2	52	26	61.6	30.8	68.8	34.4
R16	36.4	18.2	54	27	72.4	36.2	84.4	42.2	92	46
Mean	28.88	14.44	43.48	21.74	59.8	29.9	73.95	36.98	84.75	42.38
SD	6.06	3.03	5.01	2.51	5.92	2.96	6.25	3.13	6.42	3.21
SE_c	1.52	0.76	1.25	0.63	1.48	0.74	1.56	0.78	1.61	0.8

7.2.2.2.2 MRFM analysis

As noted above, the results are further analysed using MFRM with the program FACETS (Linacre, 2013). MFRM is widely used in language testing, particularly in relation to performance assessments which utilise rater judgements. One of the key benefits of MFRM is the ability to place the various variables, or facets, which contribute to measurement of the trait of interest onto a common measurement scale (Bachman, 2004; Eckes, 2011; McNamara, 1996, McNamara & Knock, 2012). In addition to item difficulty and test-taker ability, MFRM applied through FACETS can estimate the relative severity of raters, one of the key factors affecting the reliability and accuracy of performance assessments. FACETS takes account of both the relative severity of raters and difficulty of items in the final estimates of test-taker ability. These estimates are made on the logit scale, but FACETS also provides a very useful transformation of these measures back to the metric of the rating scale employed. These estimates are referred to as Fair Averages, and represent a "fair" estimate of the rating that would be achieved when the relative severity / difficulty / ability of the variables contributing to the final judgement are taken into account (Linacre, 2014). MFRM has been applied to standard setting (e.g. Engelhard, 2000; Engelhard & Stone, 1998; Lumley, Lynch & McNamarra, 1994) and in particular has been applied to standard setting in relation to linking exams to the CEFR, for example by Dunlea, (2016), O'Sullivan (2008), Papageorgiou (2007) and Eckes (2009).

To analyse the standard-setting data from round two of the Aptis Listening test judgements, a two-facet analysis was carried out, with raters and test items as facets. For the purposes of analysis, the rater judgements, originally made in 10-point increments (0, 10, 20, etc.) by judges to represent their percentage correct estimates , were converted to a rating scale with possible ratings of 0–10 (in which a 10% probability judgement is treated as a rating of 1, a 20% judgement as 2, etc.).

FACETS also provides a useful quality assurance measure of rater consistency, the infit and outfit mean square statistics. These give an estimate of the degree of fit of the observed responses to the responses predicted by the Rasch model. The infit mean square is usually reported rather than the outfit mean square, as it focuses on "the degree of fit in the most typical responses in the matrix" and is thus less susceptible to a few unpredictable outlying responses than the outfit mean square (McNamara, 1996, p. 172). In relation to standard setting, a higher fit statistic represents misfit, or unpredictability in the data, and levels of misfit greater than 1.5 would be an indication that those raters are not rating the items in the same relative order of difficulty (Engelhard & Stone, 1998, p. 185). Misfit is usually considered more problematic than overfit, or low infit mean squares, which represent response patterns that are too predictable (Myford & Wolfe, 2004).

Although various criteria for interpreting infit results have been suggested, this study uses a commonly employed threshold of 1.5 to identify rater consistency (e.g. Lunz, Wright & Linacre, 1990; Engelhard & Stone, 1998; O'Sullivan, 2008; Eckes, 2011, O'Sullivan & Dunlea, 2015; Fairbairn & Dunlea, 2017).

It is worth noting that Myford and Wolfe (2004) suggest that fit statistics in the range of 1.5 to 2.0 may still represent useful rater responses in many low-stakes situations, and Taylor and Galaczi (2012) describe using this range for identifying problematic raters in training and standardization exercises.

A separate two-facet analysis was run for the judgement data obtained for each targeted CSE level. After each initial run, infit mean square statistics for raters were examined and raters demonstrating misfit above the 1.5 criteria were dropped from the analysis.

Table 28 shows the number of analysis runs required for each CSE level to reach the required fit criteria for raters. For each analysis run, raters showing Infit Mean Squares above 1.5 are shown with the infit statistic shown in parentheses. As can be seen, analysing the rater judgements by level allows us to identify raters who seemed to experience difficulties, or at least differences in interpretation, at particular CSE levels. All raters showed sufficient consistency in interpretation for CSE 3 on the first run, for CSE 4 by the fifth run, for CSE 5 on the second, and for CSE 6 and CSE 7 on the third run. The greatest number of raters were dropped from the analysis for CSE 4, with six raters dropped, and the fewest for CSE 3, with no raters dropped from the analysis. The final fair averages for estimating cutoffs for each level used the pool of raters who showed consistent and appropriate levels of fit across all of the runs required for that level. For CSE 3, this meant 16 raters were used, 10 raters for CSE 4, 15 raters for CSE 5, 12 raters for CSE 6, and 11 raters for CSE 7. The numbers in the final pool all fall within the optimal range of raters, 10 to 15, recommended for use with the Modified Angoff approach as noted in in the literature review in Section 5.2.

Table 28: Overview of rater misfit for Aptis Listening

Run	CSE 3	CSE 4	CSE 5	CSE 6	CSE 7
1st	N/A	R12 (2.39), R5 (1.95)	R12 (1.69)	R3 (2.76), R5, 2.40	R3 (3.24), R5 (2.19), R7 (2.01)
2nd		R8 (1.71), R13 (1.63)	N/A	R12 (1.66), R6 (1.64)	R4 (2.29), R10 (1.62)
3rd		R9 (1.75)		N/A	N/A
4th		R11 (1.89)			
5th		N/A			

Figures 28 to 32 show the FACETS rater measurement reports for Aptis Listening for CSE 3 to CSE 7 for the final analysis run for each level, in which all remaining raters fit the quality assurance criteria stated above. To calculate the cutoff for each level from the MFRM output, we follow the same procedure as for the CTT results above. As already noted, the Fair Average is on the same rating scale used for judgements. So a Fair Average of 1 equates to a judgement of 10% correct for the 100 minimally competent test-takers at that level, and a Fair Average of 5 would be 50%. The rater fair average estimate is the rater's judgement across all items but adjusted for rater severity. The mean of rater fair averages in the following table is thus our cutoff estimate for that level.

Figure 28: Rater measurement report for Aptis Listening CSE 3 (final run)

CSE3 AptisL_Ang_Round2 03/12/2017 18:38:32
Table 7.1.3 Raters Measurement Report (arranged by N).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBisI	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
72	25	2.88	2.89	-.02	.22	.29	-3.5	.29	-3.5	1.65	.97	30.9	29.1	1	R1
90	25	3.60	3.64	-.84	.21	.60	-1.5	.58	-1.7	1.43	.94	24.5	24.4	2	R2
70	25	2.80	2.80	.07	.22	1.05	.2	1.06	.3	.94	.89	29.9	29.3	3	R3
71	25	2.84	2.85	.03	.22	.93	-.1	.96	.0	1.13	.91	29.1	29.2	4	R4
109	25	4.36	4.49	-1.70	.22	1.45	1.4	1.37	1.2	.65	.83	16.8	16.2	5	R5
59	25	2.36	2.28	.63	.23	.83	-.4	.79	-.6	.84	.91	26.7	28.9	6	R6
50	25	2.00	1.77	1.14	.25	.53	-1.7	.46	-1.9	1.48	.96	28.3	26.6	7	R7
58	25	2.32	2.23	.68	.23	1.03	.1	.91	-.2	1.12	.96	32.8	28.7	8	R8
70	25	2.80	2.80	.07	.22	1.34	1.1	1.25	.9	.85	.94	36.5	29.3	9	R9
65	25	2.60	2.58	.32	.22	.90	-.2	.85	-.4	1.23	.95	29.6	29.4	10	R10
71	25	2.84	2.85	.03	.22	1.31	1.1	1.20	.7	.62	.93	24.3	29.2	11	R11
51	25	2.04	1.83	1.08	.24	1.24	.8	1.01	.1	1.05	.94	31.5	27.0	12	R12
68	25	2.72	2.71	.17	.22	.90	-.2	.91	-.2	1.25	.97	38.9	29.4	13	R13
72	25	2.88	2.89	-.02	.22	.24	-3.9	.25	-3.9	1.72	.98	31.7	29.1	14	R14
88	25	3.52	3.55	-.75	.21	1.44	1.5	1.88	2.6	.32	.87	17.1	25.2	15	R15
91	25	3.64	3.68	-.89	.21	1.12	.4	1.17	.6	.77	.88	21.1	24.0	16	R16
72.2	25.0	2.89	2.87	.00	.22	.95	-.3	.93	-.4		.93			Mean (Count: 16)	
15.2	.0	.61	.69	.73	.01	.37	1.6	.40	1.7		.04			S.D. (Population)	
15.7	.0	.63	.71	.75	.01	.38	1.7	.42	1.7		.04			S.D. (Sample)	

Model, Populn: RMSE .22 Adj (True) S.D. .69 Separation 3.10 Strata 4.46 Reliability (not inter-rater) .91
Model, Sample: RMSE .22 Adj (True) S.D. .72 Separation 3.21 Strata 4.61 Reliability (not inter-rater) .91
Model, Fixed (all same) chi-square: 166.6 d.f.: 15 significance (probability): .00
Model, Random (normal) chi-square: 13.8 d.f.: 14 significance (probability): .46
Inter-Rater agreement opportunities: 3000 Exact agreements: 843 = 28.1% Expected: 815.4 = 27.2%

Figure 29: Rater measurement report for Aptis Listening CSE 4 (final run)

CSE4 AptisL_Ang_Round2 D5 20/09/2018 10:35:22
Table 7.1.3 Raters Measurement Report (arranged by N).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBisI	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
108	25	4.32	4.04	.12	.27	.82	-.6	.86	-.4	1.26	.98	35.6	35.9	1	R1
121	25	4.84	4.73	-.84	.28	1.11	.4	1.00	.1	.90	.95	33.3	33.2	2	R2
99	25	3.96	3.63	.77	.27	.98	.0	1.05	.2	1.00	.94	29.8	33.5	3	R3
102	25	4.08	3.76	.55	.27	.84	-.5	.86	-.4	1.04	.94	32.0	34.7	4	R4
110	25	4.40	4.14	-.03	.27	1.23	.8	1.31	1.0	.72	.91	28.4	35.9	6	R6
89	25	3.56	3.27	1.52	.28	.80	-.6	.71	-1.0	1.40	.95	28.4	27.6	7	R7
116	25	4.64	4.46	-.47	.27	1.05	.2	.97	.0	1.11	.95	37.8	35.0	10	R10
109	25	4.36	4.09	.04	.27	.48	-2.2	.58	-1.6	1.37	.98	36.9	35.9	14	R14
106	25	4.24	3.94	.26	.27	1.16	.6	1.26	.9	.78	.94	30.2	35.6	15	R15
135	25	5.40	5.49	-1.93	.28	.72	-.9	.69	-.9	1.20	.96	20.4	24.0	16	R16
109.5	25.0	4.38	4.16	.00	.27	.92	-.3	.93	-.2		.95			Mean (Count: 10)	
11.9	.0	.48	.59	.89	.00	.22	.9	.23	.8		.02			S.D. (Population)	
12.6	.0	.50	.62	.94	.00	.23	.9	.24	.9		.02			S.D. (Sample)	

Model, Populn: RMSE .27 Adj (True) S.D. .85 Separation 3.11 Strata 4.48 Reliability (not inter-rater) .91
Model, Sample: RMSE .27 Adj (True) S.D. .90 Separation 3.29 Strata 4.73 Reliability (not inter-rater) .92
Model, Fixed (all same) chi-square: 101.9 d.f.: 9 significance (probability): .00
Model, Random (normal) chi-square: 8.3 d.f.: 8 significance (probability): .40
Inter-Rater agreement opportunities: 1125 Exact agreements: 352 = 31.3% Expected: 372.6 = 33.1%

Figure 30: Rater measurement report for Aptis Listening CSE 5 (final run)

CSE5 AptisL_Ang_Round2 D2 : drop misfitting raters above 2.0 20/09/2018 09:41:01
Table 7.1.3 Raters Measurement Report (arranged by N).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBisI	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
143	25	5.72	5.30	.48	.24	1.49	1.5	1.74	2.1	.33	.96	29.7	31.9	1	R1
152	25	6.08	5.67	-.05	.24	.89	-.2	.86	-.4	1.05	.94	29.3	32.7	2	R2
133	25	5.32	4.96	1.08	.24	.89	-.3	1.01	.1	.98	.92	26.7	28.5	4	R4
147	25	5.88	5.46	.25	.24	1.20	.7	1.22	.7	.69	.91	32.0	32.6	6	R6
137	25	5.48	5.09	.84	.24	.60	-1.4	.57	-1.6	1.52	.94	36.3	30.2	7	R7
166	25	6.64	6.41	-.89	.25	.67	-1.1	.64	-1.2	1.37	.96	29.3	29.4	8	R8
164	25	6.56	6.30	-.77	.25	.71	-1.0	.64	-1.2	1.43	.97	36.7	30.1	9	R9
162	25	6.48	6.18	-.65	.25	.92	-.1	.84	-.4	1.22	.94	39.0	30.8	10	R10
143	25	5.72	5.30	.48	.24	1.39	1.2	1.30	1.0	.68	.92	33.7	31.9	11	R11
166	25	6.64	6.41	-.89	.25	.54	-1.8	.55	-1.6	1.53	.98	36.0	29.4	13	R13
139	25	5.56	5.16	.72	.24	.52	-1.8	.59	-1.5	1.38	.97	29.3	30.9	14	R14
130	25	5.20	4.87	1.25	.24	.98	.0	.98	.0	1.00	.92	25.7	27.0	15	R15
181	25	7.24	7.34	-1.85	.26	.69	-1.0	.74	-.6	1.11	.96	20.3	21.5	16	R16
151.0	25.0	6.04	5.73	.00	.25	.88	-.4	.90	-.4		.94			Mean (Count: 13)	
14.9	.0	.60	.71	.90	.00	.30	1.1	.34	1.1		.02			S.D. (Population)	
15.5	.0	.62	.74	.94	.00	.31	1.1	.35	1.2		.02			S.D. (Sample)	

Model, Populn: RMSE .25 Adj (True) S.D. .87 Separation 3.52 Strata 5.02 Reliability (not inter-rater) .93
Model, Sample: RMSE .25 Adj (True) S.D. .90 Separation 3.67 Strata 5.23 Reliability (not inter-rater) .93
Model, Fixed (all same) chi-square: 169.4 d.f.: 12 significance (probability): .00
Model, Random (normal) chi-square: 11.2 d.f.: 11 significance (probability): .42
Inter-Rater agreement opportunities: 1950 Exact agreements: 606 = 31.1% Expected: 580.3 = 29.8%

Figure 31: Rater measurement report for Aptis Listening CSE 6 (final run)

CSE6 AptisL_Ang_Round2 D3: drop 2 raters above 1.5 in second run 20/09/2018 10:01:34
Table 7.1.3 Raters Measurement Report (arranged by N).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBisI	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
180	25	7.20	7.12	.50	.29	.80	-.6	.77	-.5	1.17	.95	36.7	36.5	1	R1
178	25	7.12	7.03	.67	.29	.51	-1.9	.49	-1.7	1.51	.97	36.4	36.0	2	R2
160	25	6.40	6.31	2.26	.30	1.49	1.4	2.20	2.6	.26	.87	17.8	26.9	4	R4
185	25	7.40	7.33	.07	.29	1.18	.6	.95	.0	1.08	.94	39.6	37.4	7	R7
203	25	8.12	8.40	-1.61	.32	1.12	.4	.98	.2	.86	.93	33.8	34.6	8	R8
208	25	8.32	8.75	-2.14	.33	.54	-1.5	.42	-.4	1.41	.96	31.3	32.3	9	R9
195	25	7.80	7.87	-.82	.31	1.32	1.0	1.09	.3	.72	.92	37.8	37.0	10	R10
175	25	7.00	6.92	.93	.29	.96	.0	.95	.0	.76	.95	35.3	35.0	11	R11
201	25	8.04	8.27	-1.41	.32	.57	-1.5	.47	-.8	1.50	.97	37.5	35.4	13	R13
172	25	6.88	6.80	1.19	.29	.71	-1.0	.67	-.9	1.23	.97	34.5	33.8	14	R14
154	25	6.16	6.05	2.82	.31	.83	-.4	.66	-.9	1.30	.93	20.4	22.6	15	R15
211	25	8.44	8.95	-2.48	.34	.88	-.2	.96	.3	.99	.93	31.6	30.7	16	R16
185.2	25.0	7.41	7.48	.00	.31	.91	-.3	.88	-.2		.94			Mean (Count: 12)	
17.8	.0	.71	.91	1.63	.02	.30	1.0	.45	1.1		.03			S.D. (Population)	
18.6	.0	.74	.95	1.71	.02	.32	1.1	.47	1.1		.03			S.D. (Sample)	

Model, Populn: RMSE .31 Adj (True) S.D. 1.60 Separation 5.20 Strata 7.27 Reliability (not inter-rater) .96
Model, Sample: RMSE .31 Adj (True) S.D. 1.68 Separation 5.44 Strata 7.59 Reliability (not inter-rater) .97
Model, Fixed (all same) chi-square: 318.8 d.f.: 11 significance (probability): .00
Model, Random (normal) chi-square: 10.7 d.f.: 10 significance (probability): .38
Inter-Rater agreement opportunities: 1650 Exact agreements: 540 = 32.7% Expected: 548.0 = 33.2%

Figure 32: Rater measurement report for Aptis Listening CSE 7 (final run)

CSE7 AptisL_Ang_Round2 D3 : Drop further 1 rater over 2.0 in second run 20/09/2018 10:22:13
Table 7.1.3 Raters Measurement Report (arranged by N).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBisI	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
219	25	8.76	8.96	-.91	.39	1.10	.3	.80	.0	1.06	.94	45.1	46.2	1	R1
205	25	8.20	8.21	.96	.35	.97	.0	2.15	2.4	.78	.93	36.4	41.0	2	R2
187	25	7.48	7.30	2.97	.33	1.80	2.3	1.70	2.0	.20	.83	24.0	28.1	4	R4
223	25	8.92	9.17	-1.53	.40	.79	-.5	.61	-.1	1.20	.93	46.9	46.1	6	R6
223	25	8.92	9.17	-1.53	.40	.65	-.9	.61	-.1	1.21	.95	46.2	46.1	8	R8
232	25	9.28	9.66	-3.13	.44	.50	-1.5	.30	-.1	1.49	.91	45.1	42.8	9	R9
198	25	7.92	7.84	1.77	.34	1.03	.2	.96	.0	1.05	.97	34.9	36.4	11	R11
208	25	8.32	8.37	.59	.35	1.43	1.3	1.52	1.2	.46	.90	36.4	42.7	12	R12
227	25	9.08	9.39	-2.20	.42	.41	-1.9	.30	-.2	1.51	.95	46.9	45.2	13	R13
203	25	8.12	8.10	1.20	.34	.34	-2.8	.32	-2.5	1.71	.97	40.0	39.8	14	R14
172	25	6.88	6.61	4.56	.33	1.04	.2	1.12	.4	.86	.92	21.5	17.6	15	R15
230	25	9.20	9.56	-2.75	.43	1.07	.3	.67	.2	1.05	.89	42.9	43.9	16	R16
210.6	25.0	8.42	8.53	.00	.38	.93	-.3	.92	.3		.92			Mean (Count: 12)	
17.8	.0	.71	.92	2.29	.04	.41	1.4	.57	1.2		.04			S.D. (Population)	
18.5	.0	.74	.96	2.40	.04	.42	1.4	.60	1.3		.04			S.D. (Sample)	

Model, Populn: RMSE .38 Adj (True) S.D. 2.26 Separation 5.96 Strata 8.28 Reliability (not inter-rater) .97
Model, Sample: RMSE .38 Adj (True) S.D. 2.37 Separation 6.24 Strata 8.65 Reliability (not inter-rater) .97
Model, Fixed (all same) chi-square: 458.9 d.f.: 11 significance (probability): .00
Model, Random (normal) chi-square: 10.8 d.f.: 10 significance (probability): .38
Inter-Rater agreement opportunities: 1650 Exact agreements: 641 = 38.8% Expected: 654.5 = 39.7%

Table 29 presents the collated cutoffs as Fair Average percentages, those Fair Average percentages converted into a scale score on the 0–50 Aptis reporting scale, the standard deviations, and the standard error of the cutscore estimates. As can be seen by comparing the FACETS Fair Average estimates of cutoffs (calculated after trimming raters showing larger than acceptable levels of variation in their interpretation of the CSE level in question in relation to the other raters and taking into account the differential severity of raters in the remaining pool), estimates are very close to the original cutoffs in Table 27 using the entire rating pool estimates. Indeed, the average difference is only 0.6 score points on the 0–50 reporting scale, and the largest difference is 1.2 score points. To put some criterion referenced meaning to these score differences, Table 30 presents the unrounded cutoffs for both CTT and FACETS with the corresponding CEFR level that would be reported for these scores by the Aptis test system. In all cases, the substantive meaning of the scores in terms of CEFR levels would not be altered by the minor differences between the two cutoff estimates.

Table 29: Overview of Fair Average cutoff estimates for Aptis Listening

	CSE 3		CSE 4		CSE 5		CSE 6		CSE 7	
	FA	Scale	FA	Scale	FA	Scale	FA	Scale	FA	Scale
Mean	28.7	14.4	41.6	20.8	57.3	28.7	74.8	37.4	85.3	42.7
SD	6.9	3.5	5.9	3	7.1	3.6	9.1	4.6	9.2	4.6
SE_c	0.9		0.9		0.9		1.3		1.4	

Table 30: Overview of cutoff estimates and Aptis CEFR levels

	CSE 3	CSE 4	CSE 5	CSE 6	CSE 7
CTT	14.44	21.74	29.9	36.98	42.38
FACETS	14.4	20.8	28.7	37.4	42.7
CEFR	A2 (lower)	A2+	B1	B2	B2 +

7.2.2.2.3 Internal validity

As already noted in the Methodology section, an important aspect contributing to the plausibility and defensibility of the linking recommendations that will come from this project depend on the degree to which the various strands of evidence converge on a similar interpretation; in effect, are the different perspectives gained from the multi-method approach proposed here telling a similar story regarding the relationship between the tests under consideration and the CSE? In addition to the use of multiple methods as suggested by Kane, which we have adopted as a part of this linking methodology framework, it has also been noted that replicating standard setting with the same methods and same panellists will likely result in slightly different outcomes (Cizek & Bunch, 2007; Kaftandjieva, 2010). Kaftandjieva (2010) also notes that utilising different ways of deriving the cutoff score from data derived with the same method and same participants can also generate different cutoffs. In the overview of results noted above, we have employed two different ways of estimating cutoff scores for the CSE levels: one utilising a CTT approach for all raters, and the other employing MFRM analysis to refine the rating pool and adjust for severity in the estimates of judges remaining in the pool before estimating cutoffs. The very close nature of the cutoffs derived through both approaches gives us confidence in the selection, training and execution of the standard-setting methods employed, and in the ability of participants to interpret the CSE level descriptions in a coherent and consistent way.

The FACETS analysis approach itself described above provides internal validity evidence through the evaluation of rater fit statistics. The final pool of raters used to derive the FACETS fair average judgements was selected through iterative FACETS analysis runs in which misfitting raters were removed from the analysis. The cutoffs derived from the MFRM approach were thus derived by identifying those raters with the most consistent interpretation of the CSE descriptors, according to the Infit Mean Square criterion. As already noted above, even after refining the rater pool in this way, the cutoffs were very close to those estimated from the second round judgements of all raters, and this in turn, gives us some confidence in the original selection and training of participants for the judgement task. We can also examine the inter-rater correlations for judgements from the CTT analysis of the original full pool of raters as an extra indication of inter-rater reliability. The mean of inter-rater correlations for judgements at each level are presented in Table 31. The correlation matrices used to generate these average correlations were created from the second round judgements of all 16 raters in the original Listening rating panel.

Table 31: Mean Pearson correlation coefficients for Aptis L second round judgements

CSE 3	CSE 4	CSE 5	CSE 6	CSE 7
0.85	0.87	0.85	0.80	0.73

Cizek and Bunch (2007) and Tannenbaum and Wiley (2008) also suggest examining the standard deviation of individual rater's cutscores across the first and second rounds of judgements. Ideally, we would hope for increased consensus across rounds as panellists carry out discussion, examine feedback between rounds, and build a more consistent interpretation of the meaning of the CSE levels and the degree to which individual items operationalise that meaning. This trend is generally borne out across rounds for CSE 3, 4 and 5 in Table 32 below, but is reversed slightly for CSE levels 6 and 7. The differences between all rounds is, however, very small. This is likely in part due to the deliberate methodological choice of using the Basket Method as an initial first round of CSE conceptualisation in terms of item difficulty, including discussion. This approach, as described in Section 5, was derived from previous studies (e.g. Dunlea, 2016; O'Sullivan, 2015) which have demonstrated that prefacing the Angoff Method with the Basket Method can make the Modified Angoff Method judgements more accessible to participants.

Table 32: Comparison of standard deviations of Aptis Listening cutoffs in rounds 1 and 2

	CSE 3	CSE 4	CSE 5	CSE 6	CSE 7
Round 1	3.56	2.97	3.17	3.04	3.09
Round 2	3.03	2.51	2.96	3.13	3.21

Tables 31 and 32 above, which contained the mean cutoffs and standard deviations for the CTT approach and MFRM approach respectively, also contained one final indicator of internal validity, the standard error of the cutscores or SE_c . This statistic provides an estimate of the precision and replicability of the cutoffs (Dunlea, 2016). It is referred to by several different terms in the literature, with Jaeger (1991) and Cizek and Bunch (2007) simply referring to the standard error of the mean; Tannenbaum and Wiley (2008) referring to the standard error of the judges; and Kaftandjieva (2010) and Cohen, Kane, and Crookes (1999) referring to the standard error of the cutscore. The statistic is derived with the following formula:

Formula 1 $$SE_c = s_x/\sqrt{n}$$

where $\mathrm{SE_c}$ refers to the standard error of the cutscore, $\mathrm{s_x}$ is the standard deviation of the mean of the estimates for individual judges, and n is the number of judges participating in the study. Cizek and Bunch describe the interpretation of $\mathrm{SE_c}$, in a scenario with a cutscore of 32 and $\mathrm{SE_c}$ of 1, in the following way: "if the standard-setting activity were replicated, the same procedure would result in a recommended cutscore between 31 and 33", that is between +/1$\mathrm{SE_c}$. Examining Tables 31 and 32 above, we note that in all cases $\mathrm{SE_c}$ for the cutscore on the 0–50 reporting scale is less than or just over 1 score point, indicating a relatively high level of precision in the cutoff estimates.

7.2.2.3 IELTS Listening

7.2.2.3.1 CTT results

The analysis for IELTS follows the same procedures as described above, with one difference. The calculation of cutoffs below is first estimated by transforming the percentage correct Modified Angoff estimates into raw score estimates. Raw scores on IELTS Listening and Reading range from 0–40 raw score points based on the number of correctly scored items. Cutoffs then have to be evaluated in relation to the IELTS band scores. The relationship between raw score points and band scores reported here is specific to the version of the IELTS test used in this standard setting panel. All live tests are equated so that band scores will be equivalent in meaning. The number of raw score points required to reach a certain band may differ across versions. For IELTS, it is the relationship between band scores and CSE levels which is of interest, as raw scores are not reported.

As with Aptis, we first provide an overview of the Basket Method level judgements for items in Table 33 below. The average item difficulty is thus CSE 5, and the Basket Method judgements for items range from a high of 7.25 to a low of 2.81.

Table 33: Overview of Basket Method judgements for Aptis Listening Test

Mean	SD	Max	Min
5	1.11	7.25	2.81

Table 34 presents the Modified Angoff cutscore estimates from round two judgements in terms of the original percentage correct estimate and the equivalent raw score point. As with Aptis, the cutoff is first estimated for each rater by averaging that rater's judgements across all 40 items. These cutoff estimates are then averaged vertically across all raters to derive the cutoff estimate for the test by this panel of judges.

Table 35 presents the IELTS band score closest to the raw score cutoff estimate, based on the raw-to-band-score conversion for the version of IELTS Listening used in this standard-setting panel (see Table 36 below). As IELTS bands span a range of raw score points, where the cutoff falls towards the upper end of a range covered by a band, the word high is included in parentheses. For example, the raw score cutoff estimate for CSE 4 of 16.69 would fall at the high end of the range of raw score points covered by IELTS band 5 for this version.

Table 34: IELTS Listening Angoff Round 2 judgements

	CSE 4		CSE 5		CSE 6		CSE 7		CSE 8	
Rater	**%**	**Raw score**	**%**	**Raw score**	**%**	**Raw score**	**%**	**Raw score**	**%**	**Raw score**
R1	36.5	14.6	45.75	18.3	56.25	22.5	73.5	29.4	89.25	35.7
R2	39.75	15.9	52.5	21	63.5	25.4	77.75	31.1	87.5	35
R3	40.25	16.1	51.25	20.5	63.75	25.5	75.75	30.3	88	35.2
R4	42.5	17	52.25	20.9	62.25	24.9	74.5	29.8	82	32.8
R5	51.5	20.6	61.75	24.7	71.75	28.7	82	32.8	90.25	36.1
R6	37.5	15	61	24.4	75.25	30.1	89	35.6	96	38.4
R7	38	15.2	53.75	21.5	72	28.8	89.75	35.9	98.25	39.3
R8	54	21.6	71	28.4	85.25	34.1	96.25	38.5	99.25	39.7
R9	41.75	16.7	67.5	27	85.5	34.2	97	38.8	99	39.6
R10	38	15.2	56	22.4	71.25	28.5	84.75	33.9	95.75	38.3
R11	39.75	15.9	53.75	21.5	69.25	27.7	83	33.2	91	36.4
R12	44.75	17.9	63.25	25.3	79	31.6	87.75	35.1	94.25	37.7
R13	43	17.2	63	25.2	80.75	32.3	94	37.6	99	39.6
R14	39	15.6	52.75	21.1	66	26.4	77.75	31.1	90.25	36.1
R15	39	15.6	51.25	20.5	61.5	24.6	72.75	29.1	87.75	35.1
R16	42.25	16.9	53.5	21.4	67.5	27	82.25	32.9	92	36.8
Mean	41.72	16.69	56.89	22.76	70.67	28.27	83.61	33.44	92.47	36.99
SD	4.71	1.89	6.65	2.66	8.38	3.35	7.79	3.12	4.94	1.98
SE_c		0.47		0.67		0.84		0.78		0.5

Table 35: IELTS Listening band estimates based on round 2 Modified Angoff raw score cutoffs

CSE 4	CSE 5	CSE 6	CSE 7	CSE 8
5 (high)	6	6.5 (high)	7.5	8 (high)

Table 36: IELTS Listening raw to band conversion for the version used in this panel

Raw	Band	Raw	Band
1	1	**21**	5.5
2	2	**22**	6
3	3	**23**	6
4	3	**24**	6
5	3.5	**25**	6
6	3.5	**26**	6.5
7	4	**27**	6.5
8	4	**28**	6.5
9	4	**29**	6.5
10	4.5	**30**	7
11	4.5	**31**	7
12	4.5	**32**	7
13	4.5	**33**	7.5
14	5	**34**	7.5
15	5	**35**	8
16	5	**36**	8
17	5	**37**	8.5
18	5.5	**38**	8.5
19	5.5	**39**	9
20	5.5	**40**	9

7.2.2.3.2 MFRM analysis

Table 37 presents an overview of the refinement of the rater pool through the use of the Infit Mean Square fit statistic to screen misfitting raters, as described in the MFRM results for Aptis. As can be seen, all levels reached suitable levels of rater fit by the third analysis run, and the CSE 5 level in the second analysis. The number of raters dropped for exceeding the 1.5 Infit Mean Square criterion ranged from one for the CSE 5 level to five for the CSE 4 level. As with Aptis, the number of raters remaining in the pool for each level fell well within the optimal range of 10 to 15 raters identified in the literature review in Section 5.

Table 37: Overview of rater misfit for IELTS Listening

Run	CSE 4	CSE 5	CSE 6	CSE 7	CSE 8
1st	R15 (1.51), R12 (1.55), R5 (1.57)	R10 (1.62)	R10 (1.92), R7 (1.55)	R10 (2.22), R15 (1.52)	R15 (2.20), R11 (1.76)
2nd	R7 (1.67), R6 (1.61)	N/A	R12 (1.84), R15 (1.58)	R7 (1.63)	R5 (1.53)
3rd	N/A		N/A	N/A	N/A

Figures 33 to 37 present the rater measurement reports from the final FACETS analysis run for each level in which all remaining raters met the required fit criteria.

Figure 33: Rater measurement report for IELTS Listening CSE 4 (final run)

CSE4 IELTSL_Ang_Round2 D3 20/09/2018 11:33:09
Table 7.1.3 Raters Measurement Report (arranged by N).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBisI	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
146	40	3.65	3.45	.71	.19	.78	-.9	.73	-1.1	1.27	.92	34.0	31.7	1	R1
159	40	3.97	3.84	.24	.19	.97	.0	1.07	.3	.87	.88	29.8	33.2	2	R2
161	40	4.03	3.90	.17	.19	.75	-1.1	.74	-1.1	1.22	.89	34.0	33.2	3	R3
170	40	4.25	4.18	-.14	.19	.70	-1.4	.73	-1.2	1.20	.90	32.3	32.7	4	R4
216	40	5.40	5.51	-1.70	.18	1.25	1.1	1.25	1.0	.79	.84	20.5	18.4	8	R8
167	40	4.18	4.08	-.04	.19	1.09	.4	.98	.0	1.13	.90	33.5	33.0	9	R9
152	40	3.80	3.63	.49	.19	.97	.0	.93	-.2	1.11	.91	34.0	32.6	10	R10
159	40	3.97	3.84	.24	.19	.96	-.1	.94	-.1	.98	.88	27.0	33.2	11	R11
172	40	4.30	4.24	-.21	.19	1.40	1.6	1.34	1.4	.73	.93	33.5	32.4	13	R13
156	40	3.90	3.75	.35	.19	.89	-.4	.90	-.3	1.07	.88	33.0	33.0	14	R14
169	40	4.22	4.15	-.11	.19	.77	-1.0	.90	-.3	1.07	.89	31.0	32.8	16	R16
166.1	40.0	4.15	4.05	.00	.19	.96	-.2	.96	-.2		.89			Mean (Count: 11)	
17.5	.0	.44	.52	.60	.00	.21	.9	.19	.8		.02			S.D. (Population)	
18.4	.0	.46	.54	.63	.00	.22	1.0	.20	.9		.02			S.D. (Sample)	

Model, Populn: RMSE .19 Adj (True) S.D. .57 Separation 3.05 Strata 4.40 Reliability (not inter-rater) .90
Model, Sample: RMSE .19 Adj (True) S.D. .60 Separation 3.21 Strata 4.62 Reliability (not inter-rater) .91
Model, Fixed (all same) chi-square: 114.7 d.f.: 10 significance (probability): .00
Model, Random (normal) chi-square: 9.2 d.f.: 9 significance (probability): .42
Inter-Rater agreement opportunities: 2200 Exact agreements: 685 = 31.1% Expected: 692.5 = 31.5%

Figure 34: Rater measurement report for IELTS Listening CSE 5 (final run)

CSE5 IELTSL_Ang_Round2 D2: drop one rater over 1.5 20/09/2018 11:37:26
Table 7.1.3 Raters Measurement Report (arranged by N).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBisI	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
183	40	4.57	4.51	1.24	.17	1.15	.7	1.12	.6	.79	.89	17.7	20.5	1	R1
210	40	5.25	5.17	.48	.17	.80	-.8	.83	-.7	1.14	.85	23.4	25.8	2	R2
205	40	5.13	5.04	.62	.17	.75	-1.1	.76	-1.0	1.18	.87	23.9	25.1	3	R3
209	40	5.22	5.14	.51	.17	.62	-1.9	.60	-2.0	1.41	.88	29.1	25.7	4	R4
247	40	6.18	6.18	-.52	.17	1.28	1.2	1.32	1.4	.65	.65	24.6	24.9	5	R5
244	40	6.10	6.09	-.44	.16	1.10	.5	1.09	.4	.97	.87	30.4	25.3	6	R6
215	40	5.38	5.30	.35	.17	1.48	1.9	1.50	2.0	.56	.61	24.6	26.2	7	R7
284	40	7.10	7.21	-1.56	.17	1.42	1.7	1.35	1.4	.62	.75	20.2	16.4	8	R8
270	40	6.75	6.84	-1.15	.17	.51	-2.6	.49	-2.8	1.46	.91	17.1	20.2	9	R9
215	40	5.38	5.30	.35	.17	.91	-.3	.91	-.3	1.16	.88	24.5	26.2	11	R11
253	40	6.32	6.35	-.68	.17	.99	.0	1.04	.2	1.01	.92	24.3	23.9	12	R12
252	40	6.30	6.32	-.65	.17	.86	-.6	.88	-.4	1.10	.93	25.4	24.1	13	R13
211	40	5.28	5.19	.46	.17	.48	-2.8	.47	-2.9	1.49	.90	30.5	25.9	14	R14
205	40	5.13	5.04	.62	.17	1.48	1.9	1.54	2.1	.52	.61	23.9	25.1	15	R15
214	40	5.35	5.27	.37	.17	.37	-3.7	.36	-3.8	1.66	.90	32.5	26.2	16	R16
227.8	40.0	5.69	5.66	.00	.17	.95	-.4	.95	-.4		.83			Mean (Count: 15)	
27.5	.0	.69	.75	.75	.00	.35	1.8	.36	1.8		.11			S.D. (Population)	
28.4	.0	.71	.77	.78	.00	.37	1.8	.38	1.9		.12			S.D. (Sample)	

Model, Populn: RMSE .17 Adj (True) S.D. .73 Separation 4.40 Strata 6.19 Reliability (not inter-rater) .95
Model, Sample: RMSE .17 Adj (True) S.D. .76 Separation 4.56 Strata 6.41 Reliability (not inter-rater) .95
Model, Fixed (all same) chi-square: 297.2 d.f.: 14 significance (probability): .00
Model, Random (normal) chi-square: 13.4 d.f.: 13 significance (probability): .42
Inter-Rater agreement opportunities: 4200 Exact agreements: 1042 = 24.8% Expected: 1012.6 = 24.1%

Figure 35: Rater measurement report for IELTS Listening CSE 6 (final run)

CSE6 IELTSL_Ang_Round2 D3 20/09/2018 11:54:08
Table 7.1.3 Raters Measurement Report (arranged by N).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBisI	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
225	40	5.63	5.65	1.85	.17	1.17	.8	1.18	.8	.92	.93	17.0	16.9	1	R1
254	40	6.35	6.39	.98	.18	.81	-.8	.78	-.9	1.15	.81	23.4	23.8	2	R2
255	40	6.38	6.41	.95	.18	.66	-1.6	.66	-1.6	1.31	.87	25.9	24.0	3	R3
249	40	6.22	6.27	1.13	.17	.66	-1.7	.66	-1.7	1.41	.88	26.1	22.9	4	R4
287	40	7.18	7.24	-.08	.18	1.33	1.3	1.38	1.5	.63	.67	25.5	26.1	5	R5
301	40	7.53	7.63	-.55	.19	1.40	1.6	1.37	1.5	.57	.79	21.6	24.9	6	R6
341	40	8.52	8.72	-2.13	.22	1.07	.3	1.08	.4	.84	.67	17.0	16.6	8	R8
342	40	8.55	8.74	-2.18	.22	.95	-.1	.86	-.5	1.11	.84	16.6	16.3	9	R9
277	40	6.93	6.97	.25	.18	1.19	.8	1.15	.6	.76	.86	23.4	26.2	11	R11
323	40	8.07	8.25	-1.36	.20	1.31	1.3	1.17	.8	.88	.89	21.4	21.1	13	R13
264	40	6.60	6.64	.67	.18	.54	-2.3	.59	-2.0	1.40	.89	27.3	25.3	14	R14
270	40	6.75	6.79	.48	.18	.46	-2.9	.45	-3.0	1.50	.90	28.4	25.8	16	R16
282.3	40.0	7.06	7.14	.00	.19	.96	-.3	.94	-.3		.83			Mean (Count: 12)	
35.9	.0	.90	.96	1.25	.01	.31	1.5	.30	1.5		.08			S.D. (Population)	
37.5	.0	.94	1.00	1.31	.02	.33	1.6	.32	1.5		.09			S.D. (Sample)	

Model, Populn: RMSE .19 Adj (True) S.D. 1.24 Separation 6.60 Strata 9.14 Reliability (not inter-rater) .98
Model, Sample: RMSE .19 Adj (True) S.D. 1.29 Separation 6.90 Strata 9.54 Reliability (not inter-rater) .98
Model, Fixed (all same) chi-square: 483.0 d.f.: 11 significance (probability): .00
Model, Random (normal) chi-square: 10.8 d.f.: 10 significance (probability): .38
Inter-Rater agreement opportunities: 2640 Exact agreements: 602 = 22.8% Expected: 593.7 = 22.5%

Figure 36: Rater measurement report for IELTS Listening CSE 7 (final run)

CSE7 IELTSL_Ang_Round2 D2 20/09/2018 12:02:54
Table 7.1.3 Raters Measurement Report (arranged by N).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBisI	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
294	40	7.35	7.42	1.91	.19	.94	-.1	.96	-.1	1.01	.88	24.0	22.6	1	R1
311	40	7.78	7.87	1.26	.20	.94	-.1	.97	.0	1.02	.75	22.5	27.6	2	R2
303	40	7.57	7.66	1.57	.20	.90	-.3	.82	-.7	1.14	.87	24.6	25.4	3	R3
298	40	7.45	7.53	1.76	.19	.82	-.7	.86	-.5	1.17	.79	23.8	23.9	4	R4
328	40	8.20	8.31	.58	.21	1.50	1.9	1.57	2.1	.39	.64	27.9	30.8	5	R5
356	40	8.90	9.08	-.73	.23	1.19	.8	.98	.0	.92	.80	28.5	31.6	6	R6
385	40	9.63	9.82	-2.85	.34	.85	-.3	.96	.1	.94	.68	26.7	27.6	8	R8
388	40	9.70	9.87	-3.23	.37	.71	-.7	.37	-.7	1.31	.79	27.1	26.9	9	R9
332	40	8.30	8.42	.41	.21	1.36	1.4	1.21	.9	.75	.87	29.8	31.2	11	R11
351	40	8.77	8.94	-.47	.22	1.31	1.2	1.09	.4	.80	.88	27.1	31.8	12	R12
376	40	9.40	9.62	-2.00	.28	1.26	.9	.74	-.4	1.15	.84	28.8	29.2	13	R13
311	40	7.78	7.87	1.26	.20	.61	-1.9	.66	-1.6	1.33	.86	25.6	27.6	14	R14
329	40	8.23	8.34	.54	.21	.53	-2.4	.53	-2.4	1.46	.87	28.3	30.9	16	R16
335.5	40.0	8.39	8.52	.00	.23	1.00	.0	.90	-.2		.81			Mean (Count: 13)	
31.6	.0	.79	.84	1.67	.06	.29	1.3	.29	1.1		.08			S.D. (Population)	
32.9	.0	.82	.87	1.74	.06	.30	1.3	.30	1.1		.08			S.D. (Sample)	

Model, Populn: RMSE .24 Adj (True) S.D. 1.66 Separation 6.88 Strata 9.50 Reliability (not inter-rater) .98
Model, Sample: RMSE .24 Adj (True) S.D. 1.73 Separation 7.16 Strata 9.88 Reliability (not inter-rater) .98
Model, Fixed (all same) chi-square: 479.2 d.f.: 12 significance (probability): .00
Model, Random (normal) chi-square: 11.7 d.f.: 11 significance (probability): .39
Inter-Rater agreement opportunities: 3120 Exact agreements: 827 = 26.5% Expected: 881.3 = 28.2%

Figure 37: Rater measurement report for IELTS Listening CSE 8 (final run)

CSE8 IELTSL_Ang_Round2 D2 20/09/2018 12:16:30
Table 7.1.3 Raters Measurement Report (arranged by N).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBisI	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
357	40	8.93	9.02	1.75	.27	1.19	.8	1.17	.7	.75	.91	44.6	42.4	1	R1
350	40	8.75	8.83	2.22	.26	.76	-1.1	.81	-.7	1.21	.79	35.2	37.9	2	R2
352	40	8.80	8.88	2.09	.26	1.23	1.0	1.03	.2	.91	.82	40.8	39.2	3	R3
328	40	8.20	8.10	3.59	.25	.57	-2.1	.55	-2.2	1.47	.79	22.7	23.2	4	R4
361	40	9.02	9.12	1.46	.27	1.53	1.9	1.74	2.5	.24	.64	41.2	44.8	5	R5
384	40	9.60	9.76	-.71	.36	1.00	.0	.76	-.2	.96	.74	53.3	53.3	6	R6
393	40	9.82	9.94	-2.26	.50	.98	.0	7.13	2.7	.59	.49	50.8	53.5	7	R7
397	40	9.93	9.98	-3.63	.70	.45	-1.0	.12	.1	1.30	.68	54.2	52.8	8	R8
396	40	9.90	9.97	-3.19	.63	.63	-.6	.29	.0	1.19	.66	53.7	53.0	9	R9
383	40	9.57	9.73	-.59	.35	.99	.0	.65	-.6	1.07	.80	55.8	53.2	10	R10
377	40	9.43	9.57	.08	.32	1.16	.6	1.05	.2	.99	.80	50.6	51.9	12	R12
396	40	9.90	9.97	-3.19	.63	1.21	.5	.27	.0	1.07	.60	53.8	53.0	13	R13
361	40	9.02	9.12	1.46	.27	.85	-.5	.83	-.6	1.20	.77	41.3	44.8	14	R14
368	40	9.20	9.31	.91	.29	.54	-2.1	.65	-1.3	1.32	.85	43.7	48.4	16	R16
371.6	40.0	9.29	9.38	.00	.38	.93	-.2	1.22	.1		.74			Mean (Count: 14)	
20.4	.0	.51	.54	2.23	.15	.30	1.2	1.69	1.3		.11			S.D. (Population)	
21.1	.0	.53	.56	2.32	.16	.32	1.2	1.75	1.3		.11			S.D. (Sample)	

Model, Populn: RMSE .41 Adj (True) S.D. 2.19 Separation 5.34 Strata 7.45 Reliability (not inter-rater) .97
Model, Sample: RMSE .41 Adj (True) S.D. 2.28 Separation 5.54 Strata 7.73 Reliability (not inter-rater) .97
Model, Fixed (all same) chi-square: 390.1 d.f.: 13 significance (probability): .00
Model, Random (normal) chi-square: 12.5 d.f.: 12 significance (probability): .41
Inter-Rater agreement opportunities: 3640 Exact agreements: 1668 = 45.8% Expected: 1693.5 = 46.5%

Table 38 presents an overview of the cutoff estimates for each level derived through the use of Fair Averages estimated in the FACETS analysis. Table 39 provides a snapshot comparison of the raw score cutoffs estimated through the CTT approach with all 16 raters and the MFRM approach using only the raters who met the required fit criteria. The relevant IELTS band that the raw score cutoff would fall within is also listed. As with Aptis, there is no substantive difference between the cutoffs estimated through the two different approaches.[2]

Table 38: Overview of Fair Average cutoff estimates for IELTS Listening

	CSE 4		CSE 5		CSE 6		CSE 7		CSE 8	
	FA	Raw	FA	Raw	FA	Raw	FA	Raw	FA	Raw
Mean	40.5	16.2	56.6	22.64	71.4	28.56	85.2	34.08	93.8	37.52
SD	5.2	2.08	7.5	3	9.6	3.84	8.4	3.36	5.4	2.16
SE_c		0.52		0.75		0.96		0.84		0.65

Table 39: Overview of cutoff estimates and IELTS band scores

	CSE 4	CSE 5	CSE 6	CSE 7	CSE 8
CTT	16.9	22.76	28.27	33.44	36.99
IELTS Band	5 (high)	6	6.5 (high)	7.5	8 (high)
FACETS	16.2	22.64	28.56	34.08	37.52
IELTS Band	5 (high)	6	6.5 (high)	7.5	8.5

7.2.2.3.3 Internal validity

As with Aptis, the MFRM approach itself provides internal validity evidence by refining the rating pool through the use of fit analyses before estimating cutoffs with the Fair Averages for the remaining raters. The results for the two different approaches to deriving cutoffs from the same set of round two judgements has again yielded very similar cutoff estimates in terms of both raw score points and IELTS bands.

Table 40 provides the mean inter-rater correlations for Modified Angoff judgements across all 16 raters in the panel based on their second round judgements. The correlations for IELTS Listening are somewhat lower than for Aptis, with the lowest mean correlations at the upper end of the CSE levels targeted in this study, CSE 8.

Table 41 provides a comparison of the standard deviations for cutoff estimates across individual raters in rounds one and two. As with Aptis, the differences are quite small, which once again can be interpreted as giving some support for the use of the Basket Method first to help create a common interpretation of the item features in relation to CSE levels before making Modified Angoff probability judgements. The trend also supports the interpretation that this common consensus view of the CSE levels was strengthened moving from round one to round two.

As with Aptis, the standard error of the cutscore is reported for the raw score cutoffs for both the CTT approach and the MFRM approach. Again as with Aptis, this quality assurance statistic provides evidence of the precision and replicability of the cutoff estimates, with no $\mathrm{SE_c}$ estimate exceeding 1 raw score point, indicating very little potential movement in the cutoff estimates.

[2] Although the CTT CSE 8 results lists the IELTS band as 8, while the MFRM results list IELTS 8.5 for the same level, the CTT score of 36.99 is actually only 0.01 raw score points below the threshold for IELTS 8.5.

Table 40: Mean Pearson correlation coefficients for IELTS L second round judgements

CSE 4	CSE 5	CSE 6	CSE 7	CSE 8
0.70	0.67	0.63	0.57	0.51

Table 41: Comparison of standard deviations of IELTS Listening cutoffs across all panellists in rounds 1 and 2

	CSE 4	CSE 5	CSE 6	CSE 7	CSE 8
Round 1	2.00	2.75	3.43	3.38	2.22
Round 2	1.89	2.66	3.35	3.12	1.98

7.2.3 Reading

7.2.3.1 Introduction

The methodology and analysis techniques employed for Reading replicate the approach taken for Listening. The same standard-setting methods were employed, the Basket Method first followed by two rounds of the Modified Angoff method. The data is analysed in the same way, using CTT analysis and MFRM analysis to derive cutoff estimates for each test. The schedule of activities in Appendix B provides a snapshot of the full set of procedures. Note that the intermediate reflection day between standard setting for Aptis and IELTS that was employed for Listening was not factored in for Reading. As already noted, Listening was to some extent a pilot application of the procedures employed, and after the Aptis standard setting, a day was taken to review and reflect on the procedures before proceeding to IELTS. As the feedback from participants and analysis results had been positive for Listening (see Section 7.2.6 on procedural validity for participants' feedback), a review day was deemed unnecessary and dropped from the schedule.

A total of 15 raters took part in the Reading panel, with the same panellists taking part in judgements for both Aptis and IELTS.

7.2.3.2 Aptis Reading

7.2.3.2.1 CTT results

One adaptation to the standard-setting methodology and data analysis applied to Listening was made for Aptis Reading. As described in Section 6.2, the Aptis Reading test was the subject of a revision project prior to this project, and the revised test format was used in this linking project. Only two tasks in the existing format were changed, the tasks targeting CEFR levels A2 and B1. The intersentential cohesion focus of the A2-level task was maintained, but instead of one task, two shorter texts are presented to test-takers with the sentences in each jumbled. Test-takers must re-order the sentences to create a coherent short text for each. Each text is treated and scored as a separate task. Test-takers thus get a score for Task 2 and Task 3, the two revised A2 intersentential cohesion tasks. The scores, however, are not dichotomous item scores of 0/1 for each sentence placed in the correct order. Instead, test-takers are scored based on the number of sentence pairs they correctly match, as the text is designed to flow in a straightforward, linearly cohesive pattern, with each sentenced linked in some way to the following sentence. For each text, test-takers receive a task score on a scale of 0–3 based on the number of sentence pairs they have correctly matched. If a test-taker does not place any of the sentences in an order which includes any correctly paired sentences, they would receive 0. If they place the text sentences in an order with some sentences correctly paired, they receive a score of 1, if they manage to reorder the text so that most sentence pairs are placed together, they receive a 2, and if the entire text is reordered correctly, meaning all sentence pairs are correctly matched, they receive a full task score of 3.

This has implications for the standard-setting methodology. For dichotomously scored items, the Modified Angoff approach is used in the same way as for Listening, with panellists estimating how many (what percentage) out of 100 minimally competent candidates will correctly answer each item. These estimates are then averaged to derive a cutoff score for the tests for each panellist. This approach is not appropriate for the Reading Task 2 and 3, each of which is scored on a scale of 0, 1, 2, 3 as described above. Cizek and Bunch (2007) suggest that for items marked with a partial credit or rating scale approach, the Extended Angoff method is suitable. In this, the judgement task changes and panellists instead answer the question: what score would a minimally competent candidate (or what average score would 100 minimally competent candidates) achieve on this task? For the purposes of this panel, the task was worded in the following way: how much of the text would a minimally competent candidate correctly place in the right order: none (0), some (1), most (2), all (3). Cizek and Bunch (2007) further suggest that a test may include both Modified Angoff and Extended Angoff items. To derive the cutoff for the test the following steps are carried out:

1. first calculate the mean percentage correct estimates across all dichotomously scored items
2. convert the percentage correct cutoff estimate into a raw score
3. calculate the raw score estimated for each Extended Angoff item by averaging judgements for that item across judges, and sum the results for all items rated this way
4. add the raw score cutoff estimate for all dichotomous items to the sum of score estimates for Extended Angoff items
5. the total is the raw score cutoff for the whole test.

This method was followed for the Aptis Reading test, with the Modified Angoff approach applied to 19 dichotomously scored items in Tasks 1, 4, and 5, and the Extended Angoff approach applied to Tasks 2 and 3.

As with Listening, a summary overview of the Basket Method level estimates for items in the test are presented first in Table 42. For Tasks 2 and 3 above, participants made a CSE level estimate for each possible score point for the task. For each possible score point (0, 1, 2, 3), the panellists were asked to estimate at which CSE level would a test-taker first achieve a score of (0, 1, 2 or 3) on this task?

Table 42: Overview of Basket Method judgements for Aptis Listening Test

Mean	SD	Max	Min
4.41	1.32	6.4	2.13

The cutoff estimates for round two Angoff judgements are presented in Table 43. This table presents the results for CSE 3 to demonstrate how the cutoffs were calculated by combining both the Modified Angoff and Extended Angoff judgements in the way described above. The first column contains the percentage correct cutoff estimates for the 19 dichotomously scored items across all raters, with the mean of all raters and standard deviation and standard error of the cutscore at the bottom. The next two columns contain the task score estimate by each rater for Task 2 and 3, with the means at the bottom of the table. The next column converts the Modified Angoff result into a raw score for the 19 dichotomously scored items, the next column presents the total raw score (adding the raw score cutoff out of 19 to the raw scores for Task 2 and 3), and the final column converts that score to a scale score on the Aptis 0–50 reporting scale.

Table 43: Cutoff score calculation for Aptis CSE 3 Reading

	Mean (Modified Angoff)	Task 2 mean (Extended Angoff)	Task 3 mean (Extended Angoff)	Task 1, Task 4, Task 5 Raw	Total raw score	Total scale score
R1	27.89	1	1	5.3	7.3	14.6
R2	35.26	1	1	6.7	8.7	17.4
R3	28.42	1	2	5.4	8.4	16.8
R4	33.16	1	1	6.3	8.3	16.6
R5	42.11	1	2	8	11	22
R6	38.42	1	1	7.3	9.3	18.6
R7	27.37	2	1	5.2	8.2	16.4
R8	30.53	1	1	5.8	7.8	15.6
R9	35.26	1	1	6.7	8.7	17.4
R10	36.32	1	1	6.9	8.9	17.8
R11	33.16	1	1	6.3	8.3	16.6
R12	40.53	1	2	7.7	10.7	21.4
R13	38.42	1	2	7.3	10.3	20.6
R14	32.11	1	1	6.1	8.1	16.2
R15	32.63	1	1	6.2	8.2	16.4
Mean	34.11	1.07	1.27	6.48	8.81	17.63
SD	4.37	0.25	0.44	0.83	1.04	2.07
SE_c	1.09	0.06	0.11	0.21	0.26	0.52

For the sake of brevity, Table 44 omits the preceding steps and presents the collated results for the whole test in terms of total scale scores for CSE 3, CSE 4, CSE 5, CSE 6 and CSE 7.

Table 44: Aptis Reading Round 2 judgements

	CSE 3		CSE 4		CSE 5		CSE 6		CSE 7	
Rater	**%**	**Raw score**	**%**	**Raw score**	**%**	**Raw score**	**%**	**Raw score**	**%**	**Raw score**
R1	29.2	14.6	47.2	23.6	60.4	30.2	75.6	37.8	85.2	42.6
R2	34.8	17.4	46.8	23.4	59.2	29.6	82	41	92	46
R3	33.6	16.8	50.4	25.2	62.4	31.2	85.2	42.6	92.4	46.2
R4	33.2	16.6	53.6	26.8	70	35	79.2	39.6	85.6	42.8
R5	44	22	55.2	27.6	70	35	76.8	38.4	83.2	41.6
R6	37.2	18.6	56.8	28.4	74.4	37.2	82.4	41.2	89.6	44.8
R7	32.8	16.4	68.4	34.2	88.4	44.2	95.6	47.8	100	50
R8	31.2	15.6	51.6	25.8	72	36	82.4	41.2	92.4	46.2
R9	34.8	17.4	52.4	26.2	73.6	36.8	83.6	41.8	89.2	44.6
R10	35.6	17.8	48	24	64.8	32.4	78	39	92.4	46.2
R11	33.2	16.6	48.8	24.4	60.4	30.2	72	36	84	42
R12	42.8	21.4	59.6	29.8	66.4	33.2	77.2	38.6	88	44
R13	41.2	20.6	50.8	25.4	62	31	82.8	41.4	91.6	45.8
R14	32.4	16.2	54.8	27.4	75.2	37.6	84.4	42.2	90.4	45.2
R15	32.8	16.4	50.8	25.4	67.2	33.6	77.2	38.6	85.2	42.6
Mean	35.25	17.63	53.01	26.51	68.43	34.21	80.96	40.48	89.41	44.71
SD	4.14	2.07	5.40	2.70	7.46	3.73	5.33	2.67	4.25	2.13
SE_c		0.53		0.70		0.96		0.69		0.55

7.2.3.2.2 MFRM results

The inclusion of both Modified Angoff and Extended Angoff judgements required a slightly different approach for the MFRM analysis. With Listening, the Fair Average for raters across all items in the test was treated as the whole-test cutoff estimate, in the same way as is normally calculated for Modified Angoff procedures when the whole test consists of dichotomous items. In this case however, the rater Fair Average would in fact subsume to different rating scales, the 0–10 Modified Angoff items and the 0–3 Extended Angoff items. As noted above, calculating the whole-test cutoff entails first converting the percentage correct estimate for the 19 dichotomously scored items and adding those to the task scores for Task 2 and Task 3. For MFRM with Aptis Reading, the following approach was taken. First, all ratings for all judges and all items were analysed together to create a common measurement framework, to screen misfitting judges, and to estimate overall severity for each judge on the whole test. This was a two-facet analysis, as with Listening, but included two measurement models, one for the 0–10 Modified Angoff items and one 0–3 scale for the Extended Angoff items. The analysis was run successively, with misfitting raters dropped, until a final run in which all raters demonstrated suitable fit. After this final run, an anchor file was generated and all facets were anchored at their estimates from the final run, including scale steps for the rating scales. Three separate anchored analyses were then run to generate a Fair Average for each rater on the 19 dichotomous items only, for Task 2 only, and for Task 3 only. The Fair Average equating to the rater's percentage correct estimate for the 19 dichotomously scored items was converted to a raw score out of 19 and then added to that rater's Fair Average task scores for Task 2 and Task 3, resulting in a raw score cutoff estimate generated from Fair Average Estimates for each rater from the anchored MFRM analyses.

Table 45 presents an overview of the number of analysis runs required to reach the required fit criterion for all raters. As can be seen, CSE 3 achieved this on the third analysis run. For CSE 4, 5, 6 and 7, the criterion was met on the second run. As with Listening, the Infit Mean Square result for misfitting raters dropped from subsequent runs is shown in parentheses.

Table 45: Overview of rater misfit for Aptis Reading

Run	CSE 3	CSE 4	CSE 5	CSE 6	CSE 7
1st	R7 (2.23)	R9 (1.59)	R4 (1.80)	R3 (2.32)	R11 (2.80), R3 (1.52)
2nd	R13 (1.64)	N/A	N/A	N/A	N/A
3rd	N/A				

Figures 38 to 42 show the rater measurement reports for final analysis runs. Note that the Fair Averages in these measurement tables were not those used for the final cutoff calculation. All facets and rating scale steps were anchored at the values estimated in these results. As described above, three further anchored analysis runs were generated for each level to then calculate the Fair Average for each rater for combined ratings on 19 items in Tasks 1, 4, and 5, and for task scores on Task 2 and 3 separately. For the sake of brevity, the measurement reports are presented only for the final analysis run which was the basis of anchoring.

Figure 38: Rater measurement report for Aptis Reading CSE 3 (final run)

CSE3 AptisR_Ang_Round2 D3 26/09/2018 12:41:00
Table 7.1.3 Raters Measurement Report (arranged by N).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBisI	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
55	21	2.62	2.39	1.35	.33	.45	-2.0	.50	-1.6	1.50	.97	34.9	36.9	1	R1
69	21	3.29	3.02	-.18	.33	.96	.0	.94	.0	1.01	.96	42.9	42.7	2	R2
57	21	2.71	2.47	1.13	.33	1.41	1.2	1.79	1.9	.55	.97	31.3	38.6	3	R3
65	21	3.10	2.83	.26	.33	1.01	.1	.89	-.2	1.07	.95	42.5	42.7	4	R4
83	21	3.95	3.86	-1.70	.32	.75	-.7	.73	-.8	1.29	.96	33.3	32.3	5	R5
75	21	3.57	3.34	-.84	.33	1.12	.4	1.08	.3	.92	.96	39.7	39.8	6	R6
60	21	2.86	2.60	.81	.33	1.07	.3	.88	-.2	1.10	.95	40.9	40.7	8	R8
69	21	3.29	3.02	-.18	.33	1.10	.3	1.07	.3	1.02	.95	42.1	42.7	9	R9
71	21	3.38	3.12	-.40	.33	.96	.0	1.01	.1	1.00	.97	44.0	42.1	10	R10
65	21	3.10	2.83	.26	.33	.42	-2.1	.43	-2.1	1.53	.98	48.8	42.7	11	R11
80	21	3.81	3.65	-1.38	.33	1.13	.4	1.23	.7	.79	.97	31.7	35.5	12	R12
63	21	3.00	2.73	.48	.33	.84	-.3	.89	-.2	1.11	.97	38.9	42.2	14	R14
64	21	3.05	2.78	.37	.33	.75	-.6	.71	-.8	1.23	.96	43.3	42.5	15	R15
67.4	21.0	3.21	2.97	.00	.33	.92	-.2	.94	-.2		.96			Mean (Count: 13)	
8.0	.0	.38	.42	.88	.00	.27	.9	.33	1.0		.01			S.D. (Population)	
8.4	.0	.40	.44	.91	.00	.28	1.0	.34	1.0		.01			S.D. (Sample)	

Model, Populn: RMSE .33 Adj (True) S.D. .81 Separation 2.46 Strata 3.61 Reliability (not inter-rater) .86
Model, Sample: RMSE .33 Adj (True) S.D. .85 Separation 2.57 Strata 3.77 Reliability (not inter-rater) .87
Model, Fixed (all same) chi-square: 93.4 d.f.: 12 significance (probability): .00
Model, Random (normal) chi-square: 10.8 d.f.: 11 significance (probability): .46
Inter-Rater agreement opportunities: 1638 Exact agreements: 648 = 39.6% Expected: 656.9 = 40.1%

Figure 39: Rater measurement report for Aptis Reading CSE 4 (final run)

CSE4 AptisR_Ang_Round2 D2 22/09/2018 15:01:46
Table 7.1.3 Raters Measurement Report (arranged by N).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBisI	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
82	21	3.90	3.62	1.16	.27	.63	-1.2	.64	-1.2	1.43	.97	29.3	28.8	1	R1
90	21	4.29	4.01	.59	.27	.84	-.4	.84	-.4	1.13	.95	32.2	32.8	2	R2
90	21	4.29	4.01	.59	.27	1.11	.4	1.07	.3	.91	.95	30.0	32.8	3	R3
98	21	4.67	4.41	.03	.27	1.35	1.0	1.22	.7	.69	.91	28.2	34.0	4	R4
102	21	4.86	4.63	-.25	.26	.46	-1.9	.41	-2.1	1.53	.97	37.7	33.5	5	R5
106	21	5.05	4.86	-.52	.26	1.37	1.1	1.26	.8	.75	.95	31.1	32.4	6	R6
135	21	6.43	6.74	-2.48	.26	1.39	1.1	1.67	1.6	.37	.84	17.2	15.6	7	R7
93	21	4.43	4.16	.38	.27	.90	-.1	.81	-.5	1.19	.96	36.6	33.6	8	R8
102	21	4.86	4.63	-.25	.26	1.36	1.0	1.90	2.1	.42	.96	24.9	33.5	10	R10
86	21	4.10	3.81	.88	.27	.42	-2.2	.47	-1.9	1.53	.97	34.8	31.1	11	R11
104	21	4.95	4.74	-.39	.26	.86	-.3	1.25	.7	1.04	.96	31.9	33.0	12	R12
100	21	4.76	4.52	-.11	.26	.69	-.9	.80	-.5	1.17	.96	33.7	33.9	13	R13
101	21	4.81	4.58	-.18	.26	.41	-2.2	.39	-2.2	1.59	.98	38.5	33.7	14	R14
91	21	4.33	4.06	.52	.27	.39	-2.4	.38	-2.4	1.61	.97	40.7	33.1	15	R15
98.6	21.0	4.69	4.48	.00	.26	.87	-.5	.94	-.4		.95			Mean (Count: 14)	
12.3	.0	.58	.72	.84	.00	.37	1.3	.46	1.4		.03			S.D. (Population)	
12.7	.0	.61	.75	.88	.00	.39	1.4	.48	1.5		.04			S.D. (Sample)	

Model, Populn: RMSE .26 Adj (True) S.D. .80 Separation 3.03 Strata 4.37 Reliability (not inter-rater) .90
Model, Sample: RMSE .26 Adj (True) S.D. .84 Separation 3.15 Strata 4.54 Reliability (not inter-rater) .91
Model, Fixed (all same) chi-square: 144.8 d.f.: 13 significance (probability): .00
Model, Random (normal) chi-square: 12.0 d.f.: 12 significance (probability): .45
Inter-Rater agreement opportunities: 1911 Exact agreements: 610 = 31.9% Expected: 603.3 = 31.6%

Figure 40: Rater measurement report for Aptis Reading CSE 5 (final run)

CSE5 AptisR_Ang_Round2 D2 22/09/2018 13:10:02
Table 7.1.3 Raters Measurement Report (arranged by N).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBisI	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
106	21	5.05	5.12	1.84	.31	.63	-1.3	.73	-.9	1.21	.97	23.4	27.5	1	R1
112	21	5.33	5.41	1.28	.30	.88	-.3	.85	-.4	1.32	.96	30.8	32.4	2	R2
120	21	5.71	5.83	.53	.31	1.42	1.3	1.35	1.1	.60	.96	32.6	36.8	3	R3
121	21	5.76	5.88	.43	.31	1.07	.3	1.15	.5	.95	.95	35.9	37.1	5	R5
132	21	6.29	6.48	-.74	.34	1.43	1.2	1.18	.5	.81	.95	33.0	36.1	6	R6
167	21	7.95	8.50	-5.72	.39	1.17	.5	.84	1.0	-3.58	.82	18.7	18.9	7	R7
126	21	6.00	6.15	-.07	.33	.97	.0	.88	-.2	1.10	.97	39.6	37.6	8	R8
130	21	6.19	6.37	-.51	.34	.51	-1.6	.48	-1.6	1.45	.98	38.8	36.9	9	R9
126	21	6.00	6.15	-.07	.33	1.09	.3	1.11	.4	.89	.97	36.3	37.6	10	R10
106	21	5.05	5.12	1.84	.31	.66	-1.2	.74	-.8	1.33	.96	26.4	27.5	11	R11
121	21	5.76	5.88	.43	.31	1.08	.3	1.02	.1	1.00	.96	37.7	37.1	12	R12
119	21	5.67	5.77	.63	.31	1.02	.1	1.01	.1	.89	.96	30.0	36.4	13	R13
134	21	6.38	6.59	-.98	.35	.80	-.4	.61	-.9	1.32	.97	36.6	35.0	14	R14
114	21	5.43	5.51	1.10	.30	.60	-1.5	.62	-1.5	1.47	.97	33.7	33.8	15	R15
123.9	21.0	5.90	6.06	.00	.32	.95	-.2	.90	-.2		.95			Mean (Count: 14)	
14.7	.0	.70	.81	1.80	.02	.28	.9	.24	.9		.04			S.D. (Population)	
15.2	.0	.72	.84	1.87	.02	.29	1.0	.25	.9		.04			S.D. (Sample)	

Model, Populn: RMSE .33 Adj (True) S.D. 1.77 Separation 5.43 Strata 7.58 Reliability (not inter-rater) .97
Model, Sample: RMSE .33 Adj (True) S.D. 1.84 Separation 5.64 Strata 7.86 Reliability (not inter-rater) .97
Model, Fixed (all same) chi-square: 333.2 d.f.: 13 significance (probability): .00
Model, Random (normal) chi-square: 12.5 d.f.: 12 significance (probability): .40
Inter-Rater agreement opportunities: 1911 Exact agreements: 619 = 32.4% Expected: 642.7 = 33.6%

Figure 41: Rater measurement report for Aptis Reading CSE 6 (final run)

CSE6 AptisR_Ang_Round2 D2 22/09/2018 15:44:08
Table 7.1.3 Raters Measurement Report (arranged by N).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBisI	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
135	21	6.43	6.71	1.56	.32	.56	-1.4	.55	-1.5	1.52	.96	47.6	41.1	1	R1
151	21	7.19	7.62	-.12	.33	1.38	1.0	1.94	1.6	.42	.93	39.6	47.9	2	R2
144	21	6.86	7.24	.63	.33	1.24	.7	1.69	1.6	.54	.94	42.1	47.1	4	R4
138	21	6.57	6.90	1.26	.32	.63	-1.1	.83	-.4	1.33	.96	39.9	43.6	5	R5
152	21	7.24	7.67	-.23	.33	1.24	.7	1.01	.1	.86	.94	45.8	47.7	6	R6
185	21	8.81	9.42	-5.03	.46	.81	-.4	.71	1.2	1.08	.80	30.8	30.7	7	R7
152	21	7.24	7.67	-.23	.33	.45	-1.8	.36	-1.5	1.56	.98	48.7	47.7	8	R8
155	21	7.38	7.83	-.57	.34	.83	-.3	.56	-.7	1.38	.96	48.4	46.8	9	R9
150	21	7.14	7.57	-.01	.33	1.39	1.0	1.55	1.1	.56	.94	43.6	48.0	10	R10
135	21	6.43	6.71	1.56	.32	1.39	1.1	1.35	1.0	.44	.92	35.5	41.1	11	R11
139	21	6.62	6.96	1.16	.32	.88	-.2	.82	-.4	1.16	.95	45.4	44.4	12	R12
153	21	7.29	7.73	-.34	.34	.81	-.4	.98	.1	1.12	.96	45.8	47.5	13	R13
157	21	7.48	7.94	-.80	.34	.61	-1.1	.48	-.8	1.27	.96	48.4	45.9	14	R14
139	21	6.62	6.96	1.16	.32	.83	-.4	.84	-.3	1.16	.95	45.8	44.4	15	R15
148.9	21.0	7.09	7.49	.00	.34	.93	-.2	.98	.1		.94			Mean (Count: 14)	
12.5	.0	.59	.68	1.60	.03	.32	1.0	.47	1.1		.04			S.D. (Population)	
12.9	.0	.62	.70	1.66	.04	.33	1.0	.48	1.1		.04			S.D. (Sample)	

Model, Populn: RMSE .34 Adj (True) S.D. 1.57 Separation 4.61 Strata 6.49 Reliability (not inter-rater) .96
Model, Sample: RMSE .34 Adj (True) S.D. 1.63 Separation 4.80 Strata 6.73 Reliability (not inter-rater) .96
Model, Fixed (all same) chi-square: 217.2 d.f.: 13 significance (probability): .00
Model, Random (normal) chi-square: 12.3 d.f.: 12 significance (probability): .42
Inter-Rater agreement opportunities: 1911 Exact agreements: 829 = 43.4% Expected: 851.4 = 44.6%

Figure 42: Rater measurement report for Aptis Reading CSE 7 (final run)

```
CSE7 AptisR_Ang_Round2 D2 2018/09/22 16:13:12
Table 7.1.3  Raters Measurement Report  (arranged by N).

+-------------------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair(M)|   -    Model | Infit      Outfit    |Estim.| Corr. | Exact Agree. |               |
|  Score   Count  Average Average|Measure  S.E. | MnSq ZStd  MnSq ZStd |Discrm| PtBisI| Obs %  Exp % | Nu Raters     |
|-------------------------------+---------------+----------------------+------+-------+--------------+---------------|
|   159      21     7.57   7.63 |   1.19   .36 |  .67  -.9   .69  -.8 | 1.34 |   .95 |  49.6   49.3 |  1 R1         |
|   176      21     8.38   8.82 |  -1.12   .39 | 1.07   .3  2.29  2.2 |  .71 |   .95 |  46.4   52.7 |  2 R2         |
|   160      21     7.62   7.71 |   1.06   .36 | 1.14   .4  1.24   .7 |  .75 |   .94 |  51.2   50.0 |  4 R4         |
|   154      21     7.33   7.23 |   1.81   .35 |  .63 -1.0   .59 -1.2 | 1.48 |   .93 |  46.0   45.1 |  5 R5         |
|   170      21     8.10   8.43 |   -.25   .37 | 1.48  1.2  1.51  1.2 |  .44 |   .94 |  46.8   53.9 |  6 R6         |
|   196      21     9.33   9.99 |( -8.08  1.87)|Minimum               |      |   .86 |  38.5   38.5 |  7 R7         |
|   177      21     8.43   8.88 |  -1.27   .40 |  .78  -.5   .77  -.3 | 1.20 |   .96 |  52.0   52.2 |  8 R8         |
|   169      21     8.05   8.36 |   -.11   .37 |  .88  -.1   .71  -.6 | 1.33 |   .96 |  55.2   53.9 |  9 R9         |
|   177      21     8.43   8.88 |  -1.27   .40 | 1.07   .3   .97   .0 |  .97 |   .95 |  50.8   52.2 | 10 R10        |
|   166      21     7.90   8.15 |    .29   .36 |  .94   .0   .78  -.4 | 1.18 |   .96 |  52.4   53.2 | 12 R12        |
|   175      21     8.33   8.75 |   -.97   .39 |  .99   .1   .88  -.1 | 1.05 |   .96 |  54.0   53.1 | 13 R13        |
|   172      21     8.19   8.56 |   -.53   .38 |  .99   .0   .98   .0 |  .80 |   .96 |  54.0   53.8 | 14 R14        |
|   159      21     7.57   7.63 |   1.19   .36 |  .87  -.2   .92   .0 | 1.08 |   .94 |  51.6   49.3 | 15 R15        |
|-------------------------------+---------------+----------------------+------+-------+--------------+---------------|
|  170.0    21.0    8.10   8.39 |   -.62   .49 |  .96   .0 1.03   .0  |      |   .94 |              | Mean (Count: 13)  |
|   10.6     .0     .50    .70 |   2.37   .40 |  .22   .6   .45   .9 |      |   .03 |              | S.D. (Population) |
|   11.0     .0     .52    .73 |   2.47   .41 |  .23   .6   .47  1.0 |      |   .03 |              | S.D. (Sample)     |
+-------------------------------------------------------------------------------------------------------------------+
   With extremes, Model, Populn: RMSE .63  Adj (True) S.D. 2.29  Separation 3.63  Strata 5.18  Reliability (not inter-rater) .93
   With extremes, Model, Sample: RMSE .63  Adj (True) S.D. 2.39  Separation 3.79  Strata 5.39  Reliability (not inter-rater) .93
Without extremes, Model, Populn: RMSE .37  Adj (True) S.D. .97  Separation 2.61  Strata 3.81  Reliability (not inter-rater) .87
Without extremes, Model, Sample: RMSE .37  Adj (True) S.D. 1.02  Separation 2.74  Strata 3.99  Reliability (not inter-rater) .88
With extremes, Model, Fixed (all same) chi-square:  113.6  d.f.: 12  significance (probability): .00
With extremes, Model,  Random (normal) chi-square:  8.7  d.f.: 11  significance (probability): .65
Inter-Rater agreement opportunities: 1638  Exact agreements: 817 =  49.9%  Expected:  828.2 =  50.6%
-------------------------------------------------------------------------------------------------------------------
```

Table 46 presents the overview of cutoffs generated with the Fair Average approach, with standard deviations and standard error of the cutscore estimates. Table 47 provides a comparison of CTT and MFRM cutscores, with the results, as with Listening, being very close. As with Listening, the substantive meaning of scores in terms of CEFR levels report on the Aptis test would not change regardless of which cutoff was chosen. For CSE 4, although a score of 25.76 for the MFRM analysis would fall just below the B1 threshold, it is in fact only 0.24 points below that threshold.

Table 46: Overview of Fair Average cutoff estimates for Aptis Reading

	CSE 3		CSE 4		CSE 5		CSE 6		CSE 7	
	FA	**Scale**	**FA**	**Scale**	**FA**	**Scale**	**FA**	**Scale**	**FA**	**Scale**
Mean	33.08	16.54	51.52	25.76	69.30	34.65	83.91	41.95	92.31	46.15
SD	4.01	2.00	6.91	3.46	8.24	4.12	5.48	2.74	3.97	1.98
SE_c		0.50		0.92		1.10		0.73		0.55

Table 47: Overview of cutoff estimates for Aptis CTT and MFRM estimates

	CSE 3	CSE 4	CSE 5	CSE 6	CSE 7
CTT	17.63	26.51	34.21	40.48	44.71
MFRM	16.09	25.76	34.65	41.95	46.15
CEFR level	A2 (lower)	B1 / A2+	B1+	B2 (lower)	B2+

7.2.2.3.3 Internal validity

As with Listening, the MFRM approach itself provides a powerful internal validity check by refining the rating pool to those raters demonstrating a consistent and coherent interpretation of the CSE levels. A maximum of two raters were dropped for misfit, but often with fit statistics within the 1.5–2.0 range noted earlier that is not considered degrading to measurement. All levels thus maintained a number of raters for MFRM final cutoff calculation well within the range recommended as optimal for the Modified Angoff method noted earlier (10 to 15).

As with Listening, there was very little difference between the cutoff estimates using the two different analysis methods: CTT employing the whole rater pool and MFRM trimming the rater pool for misfitting raters and adjusting for rater severity.

Table 48 below provides the mean inter-rater correlations derived from the ratings of all 15 raters in the pool for their second round judgements, and Table 49 shows the comparison of standard deviations of the cutscore estimates across all raters for round one and two judgements. The trend is for decreasing standard deviation across all levels, indicating that the discussion and feedback between rounds led to increased consensus in the CSE level interpretations. Also as with Listening, referring to the main table with cutoffs for all levels above for both CTT and MFRM approaches, the standard error of the cutscore is very low, below one scale score point or just over one scale score point, for all levels. The interpretation is the same as for Listening, with these results indicating a high level of precision and replicability of the cutscore estimates regardless of the analysis method chosen.

Table 48: Mean Pearson correlation coefficients for Aptis R second round judgements

CSE 3	CSE 4	CSE 5	CSE 6	CSE 7
0.91	0.88	0.88	0.81	0.70

Table 49: Comparison of standard deviations of Aptis Reading cutoffs in rounds 1 and 2

	CSE 3	CSE4	CSE5	CSE 6	CSE 7
Round 1	3.79	4.05	3.79	3.0	2.47
Round 2	2.07	2.70	3.73	2.67	2.13

7.2.3.3 IELTS Reading

7.2.3.3.1 CTT results

No adaptations were necessary for the IELTS Reading analysis, and so the methodology and analysis replicates closely that for IELTS Listening. IELTS Reading also has a total of 40 items, and a raw score of 40 which is then converted to an IELTS band score. As with Listening, the particular conversion of raw scores to band scores is version-specific, and is determined through equating techniques prior to the use of the test version in question.

Table 50 presents an overview of the Basket Method level estimates across the 40 items in the IELTS Reading test. The mean level is closer to CSE 6, at 5.88, with mean level estimates for each item ranging from a high of 7.27 to a low of 4.60.

Table 50: Overview of Basket Method judgements for IELTS Reading Test

Mean	SD	Max	Min
5.88	0.74	7.27	4.60

Table 51 presents the collated cutoff estimates for all 15 raters in terms of both the percentage correct and the raw score.

Table 51: IELTS Reading Angoff Round 2 judgements

	CSE 4		CSE 5		CSE 6		CSE 7		CSE 8	
Rater	**%**	**Raw score**	**%**	**Raw score**	**%**	**Raw score**	**%**	**Raw score**	**%**	**Raw score**
R1	27.25	10.9	39.75	15.9	56	22.4	72.5	29	88.25	35.3
R2	34.75	13.9	46.25	18.5	60.25	24.1	78.75	31.5	92.5	37
R3	26.25	10.5	40.5	16.2	54	21.6	70.75	28.3	89.25	35.7
R4	38.25	15.3	48.25	19.3	58.25	23.3	68	27.2	78	31.2
R5	24.75	9.9	34.75	13.9	44.75	17.9	54.5	21.8	64.5	25.8
R6	24.5	9.8	43	17.2	58.5	23.4	72.5	29	88.75	35.5
R7	40.75	16.3	55.25	22.1	65.5	26.2	75.25	30.1	87.25	34.9
R8	29.25	11.7	44.25	17.7	60.25	24.1	76	30.4	91.5	36.6
R9	34.25	13.7	49.75	19.9	64.25	25.7	73.75	29.5	87.5	35
R10	35.25	14.1	48.5	19.4	63.25	25.3	78	31.2	92	36.8
R11	37	14.8	47.25	18.9	57.5	23	70	28	81.5	32.6
R12	34	13.6	45	18	60.5	24.2	77.25	30.9	90.75	36.3
R13	39.25	15.7	48.25	19.3	57.25	22.9	82.25	32.9	85.75	34.3
R14	30.5	12.2	47.75	19.1	66.5	26.6	80.75	32.3	91.5	36.6
R15	28	11.2	39	15.6	51.25	20.5	67.25	26.9	83.5	33.4
Mean	32.27	12.91	45.17	18.07	58.53	23.41	73.17	29.27	86.17	34.47
SD	5.23	2.09	4.95	1.98	5.47	2.19	6.59	2.63	7.03	2.81
SE_c	1.35	0.54		0.51	1.41	0.57	1.70	0.68	1.81	0.73

Table 52: IELTS Reading band estimates based on round 2 Modified Angoff raw score cutoffs

CSE 4	CSE 5	CSE 6	CSE 7	CSE 8
4.5 (high)	5.5	6 (high)	7	7.5 (high)

Table 53: IELTS Reading raw to band conversion for the version used in this panel

Raw	Band	Raw	Band
1	1	21	6
2	2	22	6
3	3	23	6
4	3	24	6
5	3.5	25	6.5
6	3.5	26	6.5
7	4	27	6.5
8	4	28	6.5
9	4	29	7
10	4.5	30	7
11	4.5	31	7
12	4.5	32	7.5
13	5	33	7.5
14	5	34	7.5
15	5	35	8
16	5	36	8.5
17	5.5	37	8.5
18	5.5	38	8.5
19	5.5	39	9
20	5.5	40	9

7.2.3.3.2 MFRM analysis

The MFRM analysis for IELTS follows the same pattern as for IELTS Listening, with the Modified Angoff ratings for all 40 items being made on a 0–10 scale (representing 10% increments as with Listening, i.e. 1 = 10% of candidates estimated to answer correctly). Table 54 presents the number of analysis runs using MFRM required to refine the rater pool to the point at which all raters meet the Infit Mean Square fit criterion of 1.5 or less (the Infit Mean Square for raters dropped from subsequent runs is shown in parentheses).

Table 54: Overview of rater misfit for IELTS Reading

Run	CSE 4	CSE 5	CSE 6	CSE 7	CSE 8
1st	R3 (1.82)	R6 (1.66)	R6 (1.57)	R13 (1.82), R7 (1.52)	R2 (2.73)
2nd	N/A	N/A	R13 (1.60)	N/A	N/A
3rd			N/A		

Figures 43 to 47 show the rater measurement reports for each CSE level for the final analysis run in which all raters met the fit criterion. Following the rater measurement reports, Table 55 collates the cutoff estimates for each CSE level in terms of total percentage correct and equivalent raw score (out of 40) based on rater Fair Average estimates from the MFRM analysis. Following this, Table 56 compares the CTT and MFRM cutoffs with relevant IELTS band score for that raw score point.

Figure 43: Rater measurement report for IELTS Reading CSE 4 (final run)

CSE4 IELTSL_Ang_Round2 D2 20/09/2018 13:26:56
Table 7.1.3 Raters Measurement Report (arranged by N).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBisI	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
109	40	2.72	2.70	1.43	.28	.75	-1.1	.77	-1.0	1.23	.87	36.2	35.5	1	R1
139	40	3.47	3.40	-.66	.24	1.37	1.3	1.44	1.7	.59	.78	37.1	38.3	2	R2
153	40	3.83	3.68	-1.40	.22	1.01	.1	1.24	1.0	.76	.83	32.7	33.5	4	R4
99	40	2.47	2.45	2.21	.28	1.36	1.5	1.37	1.6	.58	.58	32.3	29.7	5	R5
98	40	2.45	2.42	2.29	.28	1.02	.1	1.04	.2	.96	.76	28.8	29.1	6	R6
163	40	4.07	3.85	-1.84	.20	1.20	.8	.98	.0	1.05	.87	35.8	29.4	7	R7
117	40	2.92	2.90	.83	.27	.87	-.4	.89	-.4	1.08	.87	39.2	38.5	8	R8
137	40	3.42	3.35	-.54	.25	1.15	.6	1.08	.4	.88	.50	38.5	38.8	9	R9
141	40	3.53	3.44	-.78	.24	.66	-1.3	.61	-1.8	1.31	.85	40.0	37.7	10	R10
148	40	3.70	3.58	-1.16	.23	.22	-4.1	.21	-4.9	1.78	.94	38.7	35.4	11	R11
136	40	3.40	3.33	-.48	.25	.98	.0	.93	-.2	1.04	.94	39.8	39.0	12	R12
157	40	3.92	3.75	-1.58	.21	1.21	.8	1.46	1.7	.63	.59	30.4	31.9	13	R13
122	40	3.05	3.02	.46	.27	1.05	.3	1.02	.1	.97	.72	38.7	39.6	14	R14
112	40	2.80	2.78	1.20	.28	.35	-3.6	.34	-3.7	1.62	.89	41.2	36.8	15	R15
130.8	40.0	3.27	3.19	.00	.25	.94	-.4	.96	-.4		.78			Mean (Count: 14)	
20.6	.0	.51	.46	1.34	.03	.33	1.7	.37	1.9		.14			S.D. (Population)	
21.4	.0	.53	.48	1.40	.03	.35	1.7	.38	2.0		.14			S.D. (Sample)	

Model, Populn: RMSE .25 Adj (True) S.D. 1.32 Separation 5.25 Strata 7.33 Reliability (not inter-rater) .96
Model, Sample: RMSE .25 Adj (True) S.D. 1.37 Separation 5.45 Strata 7.61 Reliability (not inter-rater) .97
Model, Fixed (all same) chi-square: 398.5 d.f.: 13 significance (probability): .00
Model, Random (normal) chi-square: 12.6 d.f.: 12 significance (probability): .40
Inter-Rater agreement opportunities: 3640 Exact agreements: 1324 = 36.4% Expected: 1282.0 = 35.2%

Figure 44: Rater measurement report for IELTS Reading CSE 5 (final run)

CSE5 IELTSL_Ang_Round2 D2 21/09/2018 19:19:44
Table 7.1.3 Raters Measurement Report (arranged by N).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBisI	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
159	40	3.97	3.90	1.33	.26	.98	.0	.98	.0	1.02	.87	34.8	33.9	1	R1
185	40	4.63	4.54	-.30	.24	1.20	.8	1.22	.9	.74	.81	37.1	38.3	2	R2
162	40	4.05	3.97	1.13	.26	.81	-.7	.85	-.6	1.13	.85	38.1	35.1	3	R3
193	40	4.82	4.74	-.75	.23	.91	-.2	.93	-.2	1.07	.86	36.5	36.6	4	R4
139	40	3.47	3.40	2.74	.27	1.34	1.4	1.42	1.7	.52	.56	26.0	22.4	5	R5
221	40	5.53	5.35	-2.04	.20	.86	-.5	.90	-.3	1.04	.93	26.0	25.0	7	R7
177	40	4.43	4.34	.17	.25	.79	-.8	.82	-.7	1.16	.92	37.1	38.6	8	R8
199	40	4.97	4.89	-1.07	.23	1.46	1.6	1.31	1.2	.61	.59	27.1	34.7	9	R9
194	40	4.85	4.77	-.80	.23	.75	-.9	.71	-1.2	1.22	.87	39.8	36.3	10	R10
189	40	4.72	4.64	-.53	.24	.27	-4.1	.29	-4.2	1.68	.92	40.2	37.6	11	R11
180	40	4.50	4.42	-.01	.25	.90	-.3	.85	-.5	1.15	.92	40.0	38.6	12	R12
193	40	4.82	4.74	-.75	.23	1.34	1.2	1.42	1.6	.60	.57	32.7	36.6	13	R13
191	40	4.78	4.69	-.64	.24	1.06	.3	1.10	.5	.88	.73	34.8	37.1	14	R14
156	40	3.90	3.82	1.53	.26	.31	-4.1	.31	-4.0	1.65	.90	36.0	32.5	15	R15
181.3	40.0	4.53	4.45	.00	.24	.93	-.5	.94	-.4		.81			Mean (Count: 14)	
20.4	.0	.51	.49	1.22	.02	.34	1.7	.34	1.8		.13			S.D. (Population)	
21.1	.0	.53	.51	1.26	.02	.35	1.8	.35	1.8		.14			S.D. (Sample)	

Model, Populn: RMSE .24 Adj (True) S.D. 1.19 Separation 4.93 Strata 6.90 Reliability (not inter-rater) .96
Model, Sample: RMSE .24 Adj (True) S.D. 1.24 Separation 5.12 Strata 7.16 Reliability (not inter-rater) .96
Model, Fixed (all same) chi-square: 346.0 d.f.: 13 significance (probability): .00
Model, Random (normal) chi-square: 12.5 d.f.: 12 significance (probability): .40
Inter-Rater agreement opportunities: 3640 Exact agreements: 1264 = 34.7% Expected: 1256.6 = 34.5%

Figure 45: Rater measurement report for IELTS Reading CSE 6 (final run)

CSE6 IELTSL_Ang_Round2 D3 20/09/2018 13:56:38
Table 7.1.3 Raters Measurement Report (arranged by N).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBisI	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
224	40	5.60	5.51	.54	.24	.80	-.8	.84	-.6	1.14	.93	36.5	34.9	1	R1
241	40	6.03	6.01	-.40	.23	1.09	.4	1.10	.5	.90	.84	35.4	36.0	2	R2
216	40	5.40	5.29	1.00	.24	.90	-.3	.92	-.2	1.05	.86	36.3	32.6	3	R3
233	40	5.82	5.78	.03	.24	.97	.0	.96	-.1	1.04	.85	37.3	36.2	4	R4
179	40	4.47	4.38	3.32	.26	1.38	1.5	1.53	2.1	.45	.55	16.7	14.3	5	R5
262	40	6.55	6.58	-1.50	.22	1.30	1.2	1.21	.9	.79	.93	30.2	30.2	7	R7
241	40	6.03	6.01	-.40	.23	.76	-1.0	.85	-.6	1.12	.94	36.9	36.0	8	R8
257	40	6.43	6.46	-1.25	.23	1.38	1.5	1.32	1.3	.57	.62	25.8	32.1	9	R9
253	40	6.32	6.35	-1.04	.23	.67	-1.5	.67	-1.5	1.31	.90	38.3	33.5	10	R10
230	40	5.75	5.69	.20	.24	.38	-3.5	.41	-3.3	1.50	.92	36.3	35.9	11	R11
242	40	6.05	6.04	-.46	.23	.78	-.9	.79	-.9	1.15	.87	33.3	35.9	12	R12
266	40	6.65	6.67	-1.70	.22	1.10	.4	1.11	.5	.93	.74	28.8	28.4	14	R14
205	40	5.13	5.01	1.66	.25	.57	-2.2	.51	-2.5	1.54	.88	31.3	28.0	15	R15
234.5	40.0	5.86	5.83	.00	.24	.93	-.4	.94	-.4		.83			Mean (Count: 13)	
23.5	.0	.59	.63	1.34	.01	.30	1.5	.30	1.5		.12			S.D. (Population)	
24.4	.0	.61	.66	1.40	.01	.31	1.5	.32	1.5		.12			S.D. (Sample)	

Model, Populn: RMSE .24 Adj (True) S.D. 1.32 Separation 5.59 Strata 7.79 Reliability (not inter-rater) .97
Model, Sample: RMSE .24 Adj (True) S.D. 1.38 Separation 5.83 Strata 8.10 Reliability (not inter-rater) .97
Model, Fixed (all same) chi-square: 387.9 d.f.: 12 significance (probability): .00
Model, Random (normal) chi-square: 11.7 d.f.: 11 significance (probability): .39
Inter-Rater agreement opportunities: 3120 Exact agreements: 1015 = 32.5% Expected: 993.1 = 31.8%

Figure 46: Rater measurement report for IELTS Reading CSE 7 (final run)

CSE7 IELTSL_Ang_Round2 D2 20/09/2018 14:00:11
Table 7.1.3 Raters Measurement Report (arranged by N).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBisI	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
290	40	7.25	7.30	-.05	.22	.75	-1.2	.69	-1.5	1.30	.94	40.4	34.6	1	R1
315	40	7.88	7.94	-1.32	.23	1.15	.6	1.14	.6	.86	.84	29.4	31.5	2	R2
283	40	7.07	7.13	.30	.22	1.28	1.2	1.28	1.2	.77	.78	32.9	33.7	3	R3
272	40	6.80	6.85	.84	.22	.88	-.4	.87	-.5	1.11	.84	31.0	31.0	4	R4
218	40	5.45	5.25	3.79	.25	1.02	.1	1.23	.9	.65	.55	10.8	9.8	5	R5
290	40	7.25	7.30	-.05	.22	1.15	.7	1.13	.6	.84	.82	33.1	34.6	6	R6
304	40	7.60	7.65	-.74	.23	.76	-1.1	.78	-1.0	1.28	.91	37.9	34.0	8	R8
295	40	7.38	7.42	-.29	.22	1.11	.5	1.32	1.4	.74	.60	27.1	34.7	9	R9
312	40	7.80	7.86	-1.15	.23	.83	-.7	.80	-.8	1.17	.83	33.1	32.4	10	R10
280	40	7.00	7.05	.44	.22	.93	-.2	1.03	.2	1.02	.71	25.8	33.1	11	R11
309	40	7.72	7.78	-1.00	.23	.64	-1.8	.60	-2.0	1.43	.84	35.6	33.1	12	R12
323	40	8.07	8.17	-1.77	.24	.72	-1.2	.72	-1.2	1.22	.77	26.9	28.5	14	R14
269	40	6.72	6.77	.99	.23	.83	-.7	.79	-.8	1.30	.90	32.1	30.0	15	R15
289.2	40.0	7.23	7.27	.00	.23	.93	-.3	.95	-.2		.80			Mean (Count: 13)	
26.1	.0	.65	.71	1.36	.01	.19	.9	.24	1.1		.11			S.D. (Population)	
27.2	.0	.68	.74	1.42	.01	.20	.9	.25	1.1		.11			S.D. (Sample)	

Model, Populn: RMSE .23 Adj (True) S.D. 1.34 Separation 5.85 Strata 8.14 Reliability (not inter-rater) .97
Model, Sample: RMSE .23 Adj (True) S.D. 1.40 Separation 6.10 Strata 8.47 Reliability (not inter-rater) .97
Model, Fixed (all same) chi-square: 402.9 d.f.: 12 significance (probability): .00
Model, Random (normal) chi-square: 11.7 d.f.: 11 significance (probability): .39
Inter-Rater agreement opportunities: 3120 Exact agreements: 951 = 30.5% Expected: 962.3 = 30.8%

Figure 47: Rater measurement report IELTS Reading CSE 8 (final run)

CSE8 IELTSL_Ang_Round2 D2 20/09/2018 14:11:18
Table 7.1.3 Raters Measurement Report (arranged by N).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBisI	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
353	40	8.82	8.87	-.64	.27	.94	-.2	.85	-.6	1.19	.91	41.5	40.8	1	R1
357	40	8.93	8.97	-.94	.28	1.50	1.9	1.46	1.8	.49	.72	38.8	40.8	3	R3
312	40	7.80	7.86	2.10	.25	1.24	1.0	1.24	1.0	.74	.84	22.7	23.0	4	R4
258	40	6.45	6.29	5.72	.28	1.11	.5	1.31	1.2	.67	.57	3.7	3.0	5	R5
355	40	8.88	8.92	-.79	.28	1.10	.5	1.11	.5	.86	.75	37.1	40.9	6	R6
349	40	8.73	8.78	-.34	.27	1.00	.0	.94	-.2	1.10	.86	40.6	40.4	7	R7
366	40	9.15	9.20	-1.69	.30	.84	-.6	.77	-.9	1.23	.85	38.7	38.8	8	R8
350	40	8.75	8.80	-.41	.27	.69	-1.5	.74	-1.2	1.31	.69	38.5	40.5	9	R9
368	40	9.20	9.26	-1.87	.30	.73	-1.2	.64	-1.5	1.34	.79	37.9	38.0	10	R10
326	40	8.15	8.21	1.21	.25	.71	-1.3	.68	-1.5	1.33	.74	27.1	31.1	11	R11
363	40	9.07	9.12	-1.43	.29	1.00	.0	.88	-.4	1.14	.87	39.6	39.8	12	R12
343	40	8.57	8.63	.08	.26	1.31	1.3	1.26	1.1	.67	.38	33.7	38.9	13	R13
366	40	9.15	9.20	-1.69	.30	.82	-.7	.80	-.7	1.19	.75	41.0	38.8	14	R14
334	40	8.35	8.41	.69	.26	.70	-1.4	.73	-1.2	1.29	.88	34.2	35.3	15	R15
342.9	40.0	8.57	8.61	.00	.28	.98	-.1	.96	-.2		.76			Mean (Count: 14)	
28.2	.0	.71	.75	1.94	.02	.24	1.1	.26	1.1		.14			S.D. (Population)	
29.3	.0	.73	.78	2.01	.02	.25	1.1	.27	1.2		.14			S.D. (Sample)	

Model, Populn: RMSE .28 Adj (True) S.D. 1.92 Separation 6.95 Strata 9.59 Reliability (not inter-rater) .98
Model, Sample: RMSE .28 Adj (True) S.D. 1.99 Separation 7.21 Strata 9.95 Reliability (not inter-rater) .98
Model, Fixed (all same) chi-square: 680.4 d.f.: 13 significance (probability): .00
Model, Random (normal) chi-square: 12.8 d.f.: 12 significance (probability): .39
Inter-Rater agreement opportunities: 3640 Exact agreements: 1235 = 33.9% Expected: 1274.1 = 35.0%

Table 55: Overview of Fair Average cutoff estimates for IELTS Reading

	CSE 4		CSE 5		CSE 6		CSE 7		CSE 8	
	FA	**Raw**	**FA**	**Raw**	**FA**	**Raw**	**FA**	**Raw**	**FA**	**Raw**
Mean	31.9	12.76	44.5	17.8	58.3	23.32	72.7	29.08	86.1	34.44
SD	4.6	1.84	4.9	1.96	6.3	2.52	7.1	2.84	7.5	3
SE_c		0.49		0.52		0.70		0.79		0.80

Table 56: Overview of cutoff estimates and IELTS Reading CSE levels

	CSE 4	CSE 45	CSE 6	CSE 7	CSE 8
CTT	12.91	18.07	23.41	29.27	34.47
IELTS Band	4.5 (high)	5.5	6 (high)	7	7.5 (high)
FACETS	12.76	17.8	23.32	29.08	34.44
IELTS Band	4.5 (high)	5.5	6 (high)	7	7.5 (high)

7.2.3.3.2 Internal validity

The MFRM analysis approach itself for IELTS Reading, as with Listening, contributes to the internal validity of the standard-setting results. In the case of IELTS Reading, very few raters showed misfit, with those that did falling over 1.5 but within the 1.5 to 2.0 range. Nonetheless, these raters were dropped from subsequent runs to allow for Fair Average estimates from a pool of raters demonstrating suitable fit. The difference between cutoffs from the CTT approach with all raters and the MFRM approach is extremely small, and in all cases, the relevant IELTS band interpretation for that raw score point would remain the same. It is important to reiterate that for IELTS, it is the relevant band score result for CSE level cutoff interpretations that is crucial, as raw score points are not reported and the conversion will differ from version to version. The band score interpretation in terms of test-taker ability will, however, not change due to equating of all versions to ensure consistent interpretation of band scores across test administrations.

Table 57 shows the Inter-rater correlations for all 15 raters across round two Modified Angoff judgements on all items. Table 58 compares the standard deviations cutoff estimates across all raters derived from round one and round two judgements. The trend clearly demonstrates a move towards increased consensus and common interpretation of the CSE levels across rounds. The very small variation from round one to two once again can be interpreted as adding weight to the use of the Basket Method as an initial familiarisation round before embarking on Modified Angoff judgements. Comparing the SE_c for round two CTT cutoff estimates across all judges and for the MFRM cutscores using the refined rating pool once again demonstrates very high precision and high replicability of cutoff estimates.

Table 57: Mean Pearson correlation coefficients for IELTS Reading second round judgements

CSE 4	CSE 5	CSE 6	CSE 7	CSE 8
0.59	0.61	0.61	0.53	0.51

Table 58: Comparison of standard deviations of IELTS Reading cutoffs across all panellists in rounds 1 and 2

	CSE 4	CSE 5	CSE 6	CSE 7	CSE 8
Round 1	2.74	2.34	2.50	2.91	3.12
Round 2	2.09	1.98	2.19	2.63	2.81

7.2.4 Speaking

7.2.4.1 Introduction

The standard-setting methodology employed for the productive skills is necessarily different to that employed for standard setting with Listening and Reading. The standard-setting method selected for use with the productive skills components, the Analytic Judgement Method (AJM), is described in detail in Section 5.3.4.6. Some adaptations were made to the method for application in this project, and these are also described in the methodology section.

For both Speaking and Writing, a total of 16 raters participated in the standard-setting panel. The panellists were further divided into four groups for the purpose of distributing samples for rating. However, all activities and discussion were carried out as one group with all 16 raters, and all activities took place in one room with the entire group. The schedule of activities for Speaking is presented with the schedules for all panels in Appendix B and gives a good indication of the overall timing and progression of standard-setting activities employed in this application of an adapted AJM procedure.

For Speaking, a total of 70 whole-test test performances were selected from live test performances for each test. As described in the methodology section, each group rated a common set of 10 samples first before proceeding to rate a unique set of 15 samples. All samples were collated as sound files on an individual laptop and panellists accessed the sound files on their individual laptop and listened to the performances using headphones. This allowed panellists to listen at their own pace.

The AJM procedure as applied in this context is in fact very similar to a typical performance test rating scenario, with raters evaluating the proficiency level of a performance sample using a rating scale. As such, the analysis of the standard-setting results is, in some ways, more straightforward than the receptive skills. As already noted, MFRM through the FACETS program has made a significant contribution to the field of language testing particularly in the realm of performance assessments and is now commonly used as a quality assurance measure to identify misfitting raters and also in the final estimation of test-taker scores because of the benefit of taking into account differential rater severity in the final estimation of test-taker ability measures. In the application of the AJM in this project, the performance samples are the equivalent of test-takers in a typical rating scenario. We have thus opted to use MFRM analysis through FACETS as the primary analysis approach in determining the CSE level of each performance through the ratings provided by panellists.

The following sections will present the analysis and results for both Aptis and IELTS Speaking components.

7.2.4.2 Aptis Speaking

7.2.4.2.1 Analysis procedure using MFRM

The AJM procedure relies on allocating performances to a particular CSE level. The original ratings (0–50 scale scores for Aptis) are then collated for performances allocated to each level, one of several analysis approaches used to determine cutoffs for each level. MFRM analysis using the program FACETS (Linacre, 3.71) plays a crucial role in both the allocation of performances to levels and in the quality assurance to ensure internal validity of the cutoff estimates.

For Speaking, all of the panellists' ratings were collated into a concurrent data matrix for analysis, with ratings on the 10 common performances providing the means of linking all raters and all performances within a common measurement framework. FACETS is robust at handling missing data in such concurrent data matrices, and only takes into account valid responses in the estimates of the final measures. Each of the common samples would thus have 16 ratings, one for each rater, while each performance in a unique set of 15 performances will have four ratings, one for each of the raters in the group which that particular unique set of performances was allocated to.

The rating scale used is shown in Table 59. Panellists were instructed to first identify the overall CSE level which best characterised the features of the whole performance (i.e. the test-takers' responses to all tasks in the Aptis Speaking test). As with Listening and Reading, panellists were constantly reminded to refer back to the descriptors in the CSE levels to identify criterial features in the descriptors which best capture the features of the performance they were rating at the time. As with the Basket Method, panellists were told to answer the question of at which CSE level would a test-taker first produce a test performance like this? After identifying an overall CSE level, panellists were asked to make a further refinement of their decision, identifying whether the performance was at the minimally competent level, a solid performance at that level, or a high performance close to the threshold of the next higher level. For example, if a performance was first identified as representing the features relevant to a performance likely to be produced by a test-taker with a CSE 4 level of proficiency, a rating of 7–9 on the scale below, the panellists would then refine that judgement into a CSE 4 Low (rating of 7), CSE 4 Mid (rating of 8) or CSE 4 High (rating of 9). Breaking the broad CSE level allocations down into finer-grained sub-levels allows for the application of several different analysis procedures in setting potential cutoffs, which will be described further below.

Table 59: Rating scale used for Aptis Speaking samples

Rating	Level	Rating	Level	Rating	Level	Rating	Level
1	Below 2	7	4 Low	13	6 Low	19	CSE 8
2	CSE 2	8	4 Mid	14	6 Mid		
3	2 High	9	4 High	15	6 High		
4	3 Low	10	5 Low	16	7 Low		
5	3 Mid	11	5 Mid	17	7 Mid		
6	3 High	12	5 High	18	7 High		

The ratings from all panellists were first analysed with FACETS and the Infit Mean Square fit statistics examined for raters. The intention, as with the use of FACETS for analysis of rater judgements in performance tests, was to identify those raters demonstrating a consistent and coherent interpretation of the CSE levels in their ratings. As with the Listening and Reading results described above, a criterion level of 1.5 was set as the standard for fit analysis. Judges showing misfit (results higher than 1.5) were dropped and the analysis rerun. Once all raters remaining in the analysis showed sufficient fit, the Fair Average of ratings for each performance was used to allocate performances to a CSE level. As described above in the description of the use of FACETS with Listening and Reading, the Fair Averages takes into account differential rater severity and item difficulty in the final estimation on the logit scale of the proficiency measure for the performance. The Fair Average converts the logit value estimated for each performance within the common measurement framework back to the rating scale metric used by judges, i.e. 1–19.

7.2.4.2.2 Results

The panellists' response data was analysed with a two-facet analysis, with raters and performances as facets, using the 1–19 rating scale model shown in Table 59 above. Figure 48 shows the Rater Measurement report for the second analysis run in which all remaining raters showed sufficient fit. Rater 7 was dropped from the analysis after the first run due to an Infit Mean Square of 1.7 (which although is still within the range of 1.5–2.0 considered not degrading to measurement, is above the 1.5 criteria set for this study).

Figure 49 shows the facet map which places all of the elements in the analysis onto a common measurement scale measured in logits. Raters in this analysis were centred with a mean of 0 logits, and the performances, i.e. test-takers, were left un-centred, or performance, measurement report with the Fair Average reported for each performance. In this analysis, raters are negatively oriented, meaning that the higher the measure, the greater the severity. The performances are positively oriented (as test-takers would normally be in a typical performance testing scenario) meaning that the higher the measure the greater degree of proficiency, or ability, is being demonstrated. The rating scale used is shown to the left of the map.

Figure 48: Rater measurement report for Aptis Speaking final analysis run

CSE_S_Aptis_Round2_D2 23/09/2018 13:11:04
Table 7.1.3 Raters Measurement Report (arranged by N).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBis	Exact Agree. Obs %	Exact Agree. Exp %	Group	Nu	Raters
208	25	8.32	7.11	-.38	.15	.34	-2.9	.50	-1.9	1.38	.98	28.6	23.5	1	1	R1
210	25	8.40	7.19	-.43	.15	.54	-1.7	.53	-1.7	1.60	.97	25.4	23.2	1	2	R2
172	25	6.88	5.57	.47	.16	1.06	.3	.86	-.3	1.30	.96	24.9	24.0	1	3	R3
193	25	7.72	6.52	-.04	.15	.95	.0	.88	-.2	1.03	.94	30.8	24.9	1	4	R4
157	25	6.28	5.97	.26	.16	.79	-.6	.68	-1.0	1.31	.90	26.8	23.7	2	5	R5
156	25	6.24	5.92	.28	.16	.81	-.5	.71	-.9	1.18	.95	28.6	23.7	2	6	R6
108	23	4.70	3.80	1.51	.18	1.02	.1	1.16	.5	.85	.93	16.9	17.0	2	8	R8
208	25	8.32	7.04	-.34	.15	1.06	.2	.98	.0	1.00	.96	24.9	23.6	3	9	R9
221	25	8.84	7.54	-.63	.15	1.32	1.0	1.45	1.4	.50	.94	20.0	21.8	3	10	R10
224	25	8.96	7.65	-.70	.15	.70	-1.0	.63	-1.3	1.28	.96	16.8	21.2	3	11	R11
201	25	8.04	6.77	-.18	.15	.88	-.3	.85	-.4	1.27	.96	24.9	24.0	3	12	R12
212	25	8.48	7.35	-.53	.14	1.10	.4	1.33	1.1	.90	.92	12.4	19.6	4	13	R13
216	25	8.64	7.49	-.61	.14	.56	-1.7	.56	-1.8	1.25	.96	20.5	18.9	4	14	R14
154	25	6.16	5.02	.78	.16	.42	-2.2	.44	-2.1	1.46	.97	22.2	20.1	4	15	R15
163	25	6.52	5.42	.55	.16	.89	-.2	.84	-.4	1.11	.94	30.3	21.3	4	16	R16
186.9	24.9	7.50	6.42	.00	.15	.83	-.6	.83	-.6		.95				Mean (Count: 15)	
32.3	.5	1.23	1.08	.61	.01	.27	1.1	.29	1.1		.02				S.D. (Population)	
33.4	.5	1.28	1.11	.63	.01	.28	1.1	.30	1.1		.02				S.D. (Sample)	

Model, Populn: RMSE .15 Adj (True) S.D. .59 Separation 3.81 Strata 5.41 Reliability (not inter-rater) .94
Model, Sample: RMSE .15 Adj (True) S.D. .61 Separation 3.95 Strata 5.60 Reliability (not inter-rater) .94
Model, Fixed (all same) chi-square: 210.8 d.f.: 14 significance (probability): .00
Model, Random (normal) chi-square: 13.1 d.f.: 13 significance (probability): .44
Inter-Rater agreement opportunities: 1361 Exact agreements: 321 = 23.6% Expected: 299.9 = 22.0%

Figure 49: Facet map for Aptis Speaking (final run)

CSE_S_Aptis_Round2_D2 26/09/2018 18:10:43
Table 6.0 All Facet Vertical "Rulers".

Vertical = (1A,2N,SL) Yardstick (columns lines low high extreme)= 0,3,-7,4,End

Measr	-Raters	+Performances	CSEAPT
4		15	(19)
		43	17
3		6 62	16
		4 28 47	15
2		49	
	R8	19 60	14
		34 70	13
1		37	12
	R15 R16	14 22 25 48	11
	R3 R5 R6	54 64	
* 0	R4	20 39 50 57 69	10 *
	R1 R12 R2 R9	7 8 66	9
	R10 R11 R13 R14	1 17 23 32	8
-1		52 55	
		18 46	7
		5 26	6
-2		21 36 51	
		2 61 65 67	5
		3 11 29	
-3		35 53 56 59	4
		13 45	
		30 63	3
-4		38	
		31 40 42 58	2
		24 68	CSE2
-5		9 16 27	
		10 33	
-6			
		12	
		41 44	(1)
-7			Below
Measr	-Raters	+Performances	CSEAPT

Table 60 shows the Fair Average and CSE level estimation for all 70 Aptis Speaking performances. The count column is the count of ratings provided for that performance. The level column is the rounded Fair Average, representing the level on the 19-level scale used.

Table 60: Fair average and level estimation for Aptis Speaking performances

Order	Count	Fair Avg	Level	Aptis Scale	Order	Count	Fair Avg	Level	Aptis Scale
1	15	8.2	8	40	**36**	3	5.87	6	36
2	15	5.11	5	31	**37**	3	12.46	12	40
3	15	4.24	4	21	**38**	3	2.61	3	14
4	15	15.14	15	45	**39**	3	9.85	10	45
5	15	6.29	6	24	**40**	3	2.07	2	26
6	15	15.69	16	47	**41**	4	1.17	1	12
7	15	8.63	9	34	**42**	4	2.01	2	26
8	15	8.75	9	41	**43**	4	17.34	17	50
9	15	1.61	2	28	**44**	4	1.17	1	16
10	15	1.43	1	16	**45**	4	3.28	3	19
11	4	4.35	4	19	**46**	4	6.71	7	29
12	4	1.22	1	12	**47**	4	15.55	16	45
13	4	3.35	3	17	**48**	4	11	11	34
14	4	11.26	11	29	**49**	4	14.76	15	48
15	4	18.92	19	50	**50**	4	9.68	10	38
16	4	1.67	2	14	**51**	4	5.67	6	22
17	4	8.23	8	36	**52**	4	7.59	8	40
18	4	6.86	7	22	**53**	4	4	4	17
19	4	13.8	14	47	**54**	4	10.73	11	43
20	4	9.17	9	38	**55**	4	7.59	8	31
21	4	5.58	6	28	**56**	4	4.07	4	21
22	4	11	11	41	**57**	4	9.74	10	33
23	4	7.79	8	33	**58**	4	2.16	2	12
24	4	1.89	2	10	**59**	4	3.82	4	24
25	4	11	11	43	**60**	4	14.16	14	47
26	2	6.02	6	33	**61**	4	4.8	5	19
27	2	1.63	2	17	**62**	4	15.89	16	48
28	3	15.48	15	50	**63**	4	2.98	3	22
29	3	4.37	4	24	**64**	4	10.62	11	41
30	3	3.11	3	28	**65**	4	4.8	5	29
31	3	2.07	2	10	**66**	4	8.89	9	38
32	3	8.24	8	31	**67**	4	5.05	5	26
33	3	1.54	2	16	**68**	4	1.92	2	14
34	3	13.11	13	48	**69**	4	9.44	9	34
35	3	3.98	4	21	**70**	4	13.45	13	43

After using the MFRM analysis to allocate each performance to a level, the mean and median Aptis scale score was calculated for all of the performances allocated to each level, as shown in Table 61. Cutoff estimates were derived in the following ways. For the borderline method, performances allocated to the high and low categories of adjacent levels were pooled into one category; the mean of this pooled category was then taken as the cutoff estimate for the upper level. For example, all performances in the CSE 4 High and CSE 5 Low groups would be pooled and the mean scale score of this pooled group of performances would be treated as the cutoff for CSE 5. While the borderline category approach has the benefit of homing in on performances around the thresholds between levels, it has the disadvantage of reducing the sample considerably and ignoring information from all of the other performances at that level, and so an addition approach was adopted which would use all performances. In this approach, all sub-levels (e.g. CSE 4 Low, Mid, High) are collapsed so that the scale scores for all performances at that overall level (in this case, CSE 4) are taken into account. In this collapsed categories approach, the cutoff estimate is derived by taking either the mean of the means of two adjacent levels, or the midpoint between the medians of two adjacent levels. Table 62 presents the cutoffs for each level targeted in the Aptis study, CSE 3 to CSE 7, for the three methods described above.

Table 61: Mean and median Aptis scale scores for each CSE level

Level	Mean	Median
CSE 2	18.2	17
CSE 3	24.7	24
CSE 4	34.4	34
CSE 5	38.7	40.5
CSE 6	46.9	47
CSE 7	47.5	47.5
CSE 8		

Table 62: Cutoff estimates for CSE 3 to CSE 7 for Aptis Speaking

Borderline		Collapsed categories	
Level		Mean of mean	Midpoint of median
CSE 3	20.6	21.45	20.5
CSE 4	27.7	29.55	29
CSE 5	37.6	36.55	37.25
CSE 6	43.7	42.8	43.75
CSE 7	47.2	47.2	47.25

7.2.4.2.3 Internal validity.

The different methodology and analysis procedures employed for the AJM mean that the same approaches to demonstrating internal validity of the cutoffs from standard setting used in the Listening and Reading are not applicable. The standard error of the cutscore is not applied here, for instance, as individual cutoffs for each rater are not estimated. In the approach applied here, we have instead approached the analysis in the same fashion as a performance test rating exercise.

From that perspective, we have maximised information from a pool of raters, 16 for the common performance and four for each of the samples unique to each group. In this way, the robustness of the final level estimates is being strengthened in the same way as would be the case for a speaking test rating scenario through the use of multiple raters. A typical measure of the quality of ratings in a performance testing situation is the inter-rater correlations. As one additional indicator of internal validity, a correlation matrix was thus generated for the round 2 ratings for the 10 common performance samples rated by all 16 panellists, and the mean inter-rater correlation calculated, which was 0.84.

As already noted, however, the primary focus on quality assurance in this analysis has been provided by the use of MFRM analysis through the FACETS program. MFRM has been employed in the fashion that has become standard practice in performance assessments to both refine the rating pool through the use of fit statistics to ensure that raters with a consistent and shared interpretation of the scale are used in the final estimation of measures. In addition, the final level estimation is based on the MFRM fair average, which as described above takes into account differential rater severity in the final estimate of the CSE level to which each performance is allocated.

7.2.4.3 IELTS Speaking

7.2.4.3.1 Analysis procedure using MFRM

The same analysis approach as described for Aptis was employed for IELTS. The 16 raters rated a common group of 10 samples to provide sufficient linking to ensure all ratings could be analysed within a common measurement framework and measures placed onto a common scale. The rating scale employed was slightly modified to reflect the CSE levels targeted for the IELTS test in this project, CSE 4 to CSE 8. Table 63 shows the rating scale with the CSE levels, split into low, mid and high sublevels.

Table 63: Rating scale used for IELTS Speaking samples

Rating	Level	Rating	Level	Rating	Level	Rating	Level
1	Below 3	7	5 Low	13	7 Low	19	CSE 9
2	CSE 3	8	5 Mid	14	7 Mid		
3	3 High	9	5 High	15	7 High		
4	4 Low	10	6 Low	16	8 Low		
5	4 Mid	11	6 Mid	17	8 Mid		
6	4 High	12	6 High	18	8 High		

7.2.4.3.2 Results

One rater was dropped from the first analysis run, R10 with an Infit Mean Square of 1.85. Figure 50 shows the rater measurement report from the second analysis run in which all remaining 15 raters showed acceptable levels of fit. Figure 51 shows the facet map, which can be interpreted in the same way as described for Aptis above.

Figure 50: Rater measurement report for IELTS Speaking final analysis run

CSE_S_IELTS_Round2_D2 23/09/2018 14:08:03
Table 7.1.3 Raters Measurement Report (arranged by N).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBis	Exact Agree. Obs %	Exact Agree. Exp %	Group	Nu	Raters
223	23	9.70	8.19	.11	.13	.66	-1.1	.70	-1.0	1.14	.92	19.0	19.9	1	1	R1
235	25	9.40	7.98	.18	.12	1.26	.9	1.30	1.0	.77	.85	16.4	19.6	1	2	R2
244	25	9.76	8.36	.05	.12	.96	.0	.94	-.1	1.10	.93	23.0	19.9	1	3	R3
252	25	10.08	8.70	-.06	.12	.44	-2.3	.48	-2.1	1.50	.95	22.4	19.9	1	4	R4
231	25	9.24	7.93	.20	.12	.90	-.2	.88	-.3	1.20	.91	21.1	19.7	2	5	R5
223	25	8.92	7.58	.32	.12	1.22	.8	1.07	.3	.96	.92	17.8	19.0	2	6	R6
268	25	10.72	9.56	-.35	.12	.78	-.7	.85	-.4	1.17	.92	15.7	18.6	2	7	R7
230	25	9.20	7.89	.22	.12	.66	-1.2	.64	-1.3	1.28	.93	16.8	19.6	2	8	R8
232	25	9.28	9.28	-.26	.12	1.05	.2	1.12	.4	.76	.85	15.9	19.7	3	9	R9
235	25	9.40	9.41	-.30	.12	.67	-1.2	.74	-.8	1.22	.90	21.8	19.5	3	11	R11
218	25	8.72	8.67	-.05	.12	.58	-1.5	.62	-1.4	1.48	.96	27.1	20.1	3	12	R12
236	25	9.44	8.69	-.06	.12	.86	-.3	.85	-.4	1.19	.94	16.8	19.9	4	13	R13
246	25	9.84	9.12	-.21	.12	.46	-2.2	.47	-2.1	1.49	.95	23.2	19.6	4	14	R14
202	25	8.08	7.20	.45	.12	1.22	.8	1.15	.5	.90	.86	22.7	17.5	4	15	R15
249	25	9.96	9.25	-.25	.12	.68	-1.1	.94	-.1	1.08	.94	18.9	19.4	4	16	R16
234.9	24.9	9.45	8.52	.00	.12	.83	-.6	.85	-.5		.92					Mean (Count: 15)
15.3	.5	.60	.70	.24	.00	.26	1.0	.24	.9		.03					S.D. (Population)
15.8	.5	.62	.72	.25	.00	.27	1.1	.25	1.0		.03					S.D. (Sample)

Model, Populn: RMSE .12 Adj (True) S.D. .20 Separation 1.65 Strata 2.54 Reliability (not inter-rater) .73
Model, Sample: RMSE .12 Adj (True) S.D. .21 Separation 1.73 Strata 2.64 Reliability (not inter-rater) .75
Model, Fixed (all same) chi-square: 55.8 d.f.: 14 significance (probability): .00
Model, Random (normal) chi-square: 11.2 d.f.: 13 significance (probability): .59
Inter-Rater agreement opportunities: 1359 Exact agreements: 270 = 19.9% Expected: 264.6 = 19.5%

Figure 51: Facet map for IELTS Speaking

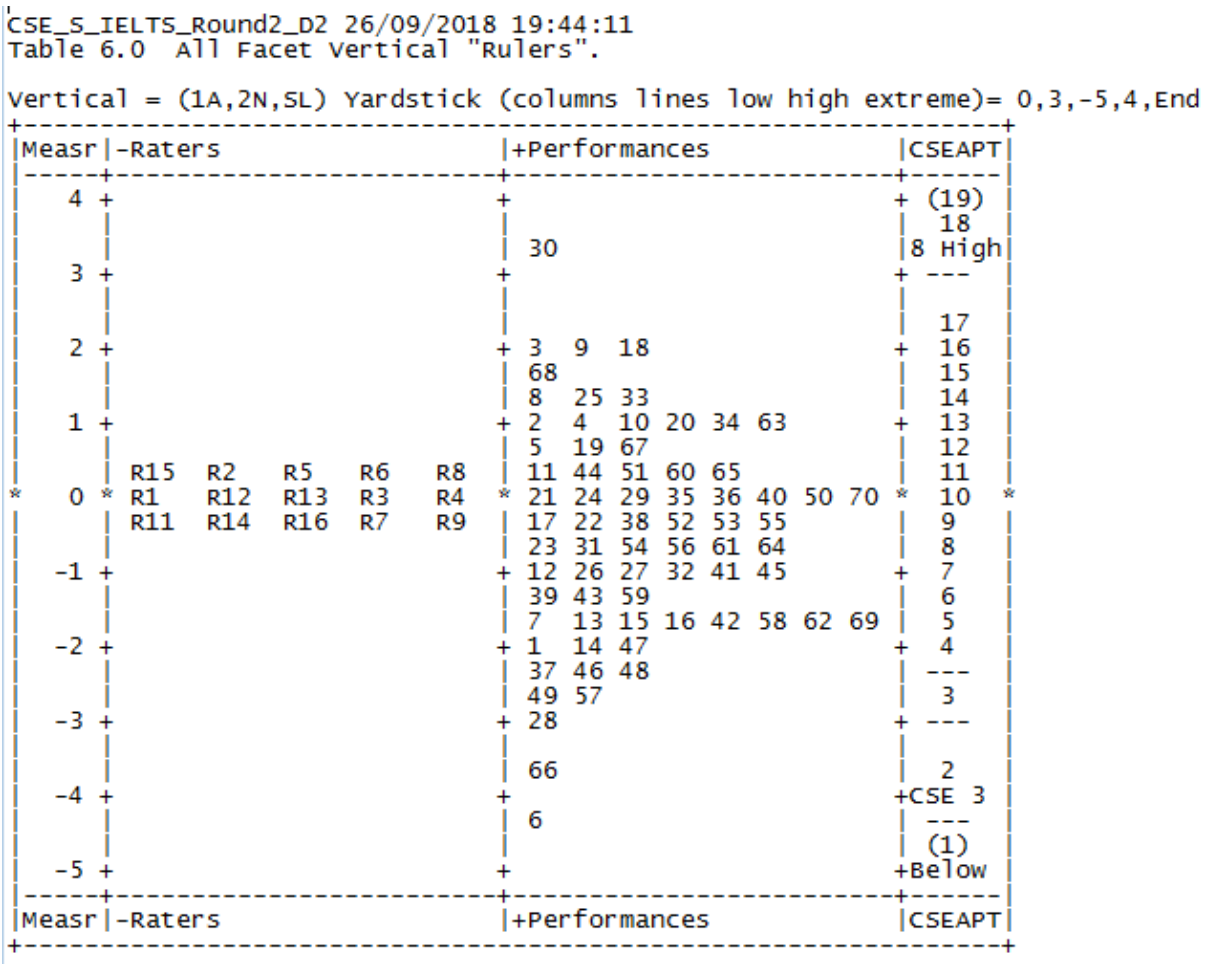


Table 64 on the following page can be read in the same way as the same table presented for Aptis results. The count is the number of ratings for each performance, the Fair Average is the MFRM measurement result taking into account rater severity transformed back to the 1–19 rating scale used by panellists to allocate each performance to a CSE overall and sub-level. The Level column is the rounded final level according to the 1–19 rating scale. Finally, for IELTS performances, the original IELTS band score allocated to the performance is shown. Three performances, 18, 66, and 68 are shaded. As can be seen from the table, the original IELTS band score and the CSE level allocated showed a considerable discrepancy for these performances. Performance 66 was very clearly allocated to quite a low level by panellists and the MFRM analysis, but had an IELTS band of 7.5. Conversely, performance 18 and 68 were both allocated to a very high level by panellists and the MFRM analysis, but had received a noticeably low IELTS band score, 4 and 3 respectively. While some variation is expected in the analysis, these significant outliers potentially would skew mean results for the level in a misleading way, given the small number of samples in each level. The result only came to light during the post-hoc analysis, and as such, there was no facility to revisit the ratings provided by panellists, or to re-rate the original IELTS band score. As such, these three performances were dropped from the final cutoff calculations presented below.

Table 65 presents the mean and median IELTS band scores for all performances allocated to that level. Table 66 presents the three approaches to deriving cutscores that were explained above for Aptis: the borderline approach using the mean of only those performances allocated to the high and low sub-levels for adjacent categories, and the mean of means and mid-point of medians for all performances allocated to whole CSE level levels (collapsing sub-levels) for adjacent levels. For the borderline method, no cutoff could be set for CSE 8, as no performances had been allocated to the CSE 7 High level.

Table 64: Fair average and level estimation for IELTS Speaking performances

Order	Count	Fair Avg	Level	IELTS Band	Order	Count	Fair Avg	Level	IELTS Band
1	15	4.01	4	4	**36**	4	10.02	10	7
2	15	13.04	13	6	**37**	4	3.68	4	5
3	15	16.51	17	8	**38**	4	9.27	9	6
4	15	12.91	13	5.5	**39**	4	6.29	6	6
5	15	12.19	12	6	**40**	4	10.02	10	5.5
6	15	1.47	1	4	**41**	3	6.76	7	6.5
7	15	5.2	5	5	**42**	3	5.47	5	5
8	15	14.23	14	5.5	**43**	3	5.78	6	5
9	15	15.83	16	8.5	**44**	3	10.64	11	7
10	15	12.71	13	8	**45**	3	6.76	7	5.5
11	3	10.5	11	5	**46**	3	3.63	4	4.5
12	3	7.2	7	5.5	**47**	3	4.25	4	8
13	4	5.2	5	5	**48**	3	3.94	4	5
14	4	4.69	5	5	**49**	3	3.01	3	4.5
15	4	5.46	5	7	**50**	3	9.68	10	7
16	4	5.2	5	4	**51**	3	10.64	11	5.5
17	4	8.96	9	8.5	**52**	3	9.05	9	7.5
18	*4*	*16.16*	*16*	*4*	**53**	3	9.36	9	6.5
19	4	12.21	12	8.5	**54**	3	8.41	8	6
20	4	13.44	13	6.5	**55**	3	9.05	9	6
21	4	9.71	10	7.5	**56**	4	7.97	8	5
22	4	8.96	9	6	**57**	4	3.22	3	4.5
23	4	8.22	8	5.5	**58**	4	5.2	5	7.5
24	4	9.46	9	5	**59**	4	6.46	6	4.5
25	4	14.19	14	7.5	**60**	4	10.66	11	7.5
26	4	7.06	7	5.5	**61**	4	8.22	8	3.5
27	4	6.8	7	5	**62**	4	5.45	5	7
28	4	2.61	3	4	**63**	4	13.18	13	6.5
29	4	10.28	10	6	**64**	4	7.73	8	4.5
30	4	17.8	18	8	**65**	4	10.66	11	5.5
31	4	7.81	8	6	***66***	*4*	*1.98*	*2*	*7.5*
32	4	7.06	7	6	**67**	4	12.44	12	7.5
33	4	14.02	14	8	***68***	*4*	*15.2*	*15*	*3*
34	4	13.27	13	6.5	**69**	4	5.45	5	6.5
35	4	10.28	10	5.5	**70**	4	10.17	10	4

Table 65: Mean and median IELTS scale scores for each CSE level

Level	Mean	Median
CSE 3	4.3	4.5
CSE 4	5.5	5
CSE 5	5.8	6
CSE 6	6.3	6
CSE 7	6.7	6.5
CSE 8	8.2	8
CSE 9	N/A	N/A

Table 66: Cutoff estimates for CSE 4 to CSE 8 for IELTS Speaking

Borderline		Collapsed categories	
Level		Mean of mean	Midpoint of median
CSE 4	4.9	4.9	4.75
CSE 5	5.5	5.65	5.5
CSE 6	6.3	6.05	6
CSE 7	6.8	6.5	6.25
CSE 8	N/A	7.45	7.25

7.2.4.3.3 Internal validity

The same comments made for Aptis above apply to the interpretation of internal validity for IELTS. Again, inter-rater correlations were calculated for all raters on the 10 common performances, and then averaged to produce a mean inter-rater correlation of 0.92. As with Aptis, however, the main approach to maximising the quality of the judgements and final cutoffs estimated from the levels performances were allocated to, is through the application of the MFRM procedure itself. This procedure has allowed us to maximise the multiple-rater design to elicit ratings on 70 performances, many more than would be possible without the ability to place the unique samples rated by each group onto the same common measurement scale. MFRM also ensures that only ratings from raters demonstrating adequate fit are included in the final analysis, and that even within that group, differential rater severity is taken into account when estimating the Fair Average for each performance.

7.2.5 Writing

7.2.5.1 Introduction

Writing followed substantially the same procedure and analysis methodology as for Speaking. The major difference is the number of samples which were able to be accommodated in the time available. For Writing, 120 whole-test performances were rated for both Aptis and IELTS. In addition, all panellists provided judgements on all performances, providing a fully crossed design. This greatly enhances the robustness of the multiple rating design generally, and ensures that all performances are linked and able to be analysed within a common measurement frame for MFRM.

7.2.5.2 Aptis Writing

7.2.5.2.1 Analysis procedure using MFRM

The same rating scale was used for Aptis Writing as was used for Aptis Speaking. The scale is shown in Table 67, indicating the relevant overall CSE level and low, mid and high sub-levels.

Table 67: Rating scale used for Aptis Writing samples

Rating	Level	Rating	Level	Rating	Level	Rating	Level
1	Below 2	7	4 Low	13	6 Low	19	CSE 8
2	CSE 2	8	4 Mid	14	6 Mid		
3	2 High	9	4 High	15	6 High		
4	3 Low	10	5 Low	16	7 Low		
5	3 Mid	11	5 Mid	17	7 Mid		
6	3 High	12	5 High	18	7 High		

As with Speaking, the ratings from all panellists were first analysed with FACETS and the Infit Mean Square fit statistics examined for raters. As with the Listening, Reading, and Speaking results described above, a criterion level of 1.5 was set as the standard for fit analysis.

7.2.5.2.2 Results

As with Speaking, the panellists' response data was analysed with a two-facet analysis, with raters and performances as facets, using the 1–19 rating scale model. For Aptis Writing, the analysis was run three times. In the first run Rater 9, with Infit Mean Square of 2.11 was dropped and in the second run, Rater 10, with an Infit Mean Square of 1.52 was dropped. Figure 52 shows the Rater Measurement report for the third analysis run in which all remaining raters showed sufficient fit. Figure 53 shows the facet map which places all of the elements in the analysis onto a common measurement scale measured in logits. The facet map is interpreted in the same way as described for Aptis and IELTS Speaking results.

Figure 52: Rater measurement report for Aptis Writing final run

CSE_W Aptis Round2 D2 23/09/2018 12:09:56
Table 7.1.3 Raters Measurement Report (arranged by N).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBis	Exact Agree. Obs %	Exact Agree. Exp %	Group	Nu	Raters
899	120	7.49	6.82	-.18	.05	.75	-1.9	.88	-.8	.99	.92	16.3	17.6	1	1	R1
887	120	7.39	6.72	-.14	.05	1.13	.9	1.01	.0	1.01	.89	19.2	17.8	1	2	R2
830	120	6.92	6.26	.01	.05	.71	-2.3	.80	-1.5	1.21	.94	18.3	18.4	1	3	R3
818	120	6.82	6.16	.05	.05	.53	-4.1	.78	-1.7	1.35	.94	21.7	18.5	1	4	R4
834	120	6.95	6.29	.00	.05	.68	-2.6	.64	-3.0	1.19	.94	18.3	18.4	2	5	R5
794	120	6.62	5.96	.12	.05	.61	-3.3	.58	-3.6	1.38	.93	20.7	18.5	2	6	R6
831	120	6.93	6.26	.01	.05	1.11	.8	1.03	.2	1.02	.91	19.9	18.4	2	7	R7
761	120	6.34	5.69	.21	.05	1.25	1.7	1.10	.7	1.07	.89	19.7	18.3	2	8	R8
760	120	6.33	5.68	.21	.05	1.06	.4	1.00	.0	1.00	.89	18.5	18.3	3	11	R11
846	120	7.05	6.39	-.03	.05	1.26	1.8	1.13	.9	.95	.90	19.8	18.3	3	12	R12
1047	120	8.73	8.05	-.57	.05	1.39	2.6	1.43	2.9	.63	.88	12.1	13.7	4	13	R13
800	120	6.67	6.01	.10	.05	1.08	.5	1.01	.1	1.09	.90	19.6	18.5	4	14	R14
742	120	6.18	5.53	.27	.05	1.15	1.1	1.14	1.0	.94	.93	17.9	18.2	4	15	R15
857	120	7.14	6.47	-.06	.05	1.20	1.4	1.36	2.4	.62	.88	14.4	18.2	4	16	R16
836.1	120.0	6.97	6.31	.00	.05	.99	-.2	.99	-.2		.91				Mean (Count: 14)	
73.3	.0	.61	.60	.20	.00	.27	2.1	.23	1.8		.02				S.D. (Population)	
76.1	.0	.63	.63	.21	.00	.28	2.2	.24	1.9		.02				S.D. (Sample)	

Model, Populn: RMSE .05 Adj (True) S.D. .19 Separation 3.68 Strata 5.24 Reliability (not inter-rater) .93
Model, Sample: RMSE .05 Adj (True) S.D. .20 Separation 3.83 Strata 5.44 Reliability (not inter-rater) .94
Model, Fixed (all same) chi-square: 208.6 d.f.: 13 significance (probability): .00
Model, Random (normal) chi-square: 12.2 d.f.: 12 significance (probability): .43
Inter-Rater agreement opportunities: 10920 Exact agreements: 1999 = 18.3% Expected: 1960.0 = 17.9%

Figure 53: Facet map for Aptis Writing

CSE_W Aptis Round2 D3 26/09/2018 21:38:34
Table 6.0 All Facet Vertical "Rulers".

Vertical = (1A,2N,SL) Yardstick (columns lines low high extreme)= 0,6.5,-4,2,End

Measr	-Raters	+Performances	CSEW_A
2		2 47 48 49 52 61	(19)

			16

			15
1		53 57 58 59 60	14
		15 54 62 64	---
		46 67	13
		56 69 93 95 108	12
		55 65 70	11
	R15	10 92	---
	R11 R14 R6 R8	9 11 24 116	10
* 0 *	R12 R16 R3 R4 R5 R7	* 25 39 50	* 9 *
	R1 R2	13 20 88	---
		71 72	8
		17 36 40 113 117	---
	R13	6 16 66 96 115	7
		18 21 30 37 74	4 Low
-1		5 7 12 33	---
		23 68 104 106 107 112	6
		14 28 29 31 38 43 85 105 109	---
		19 27 84 94 101 103 118 120	5
		44	---
		26 32 45 51 91 114	4
		3 22 35 83 111	3 Low
-2		63 82 98 119	---
			3
		34 86 87 100 102	2 High
		8 90 99	---
		4 73 75 77 81 97 110	
		79 80	2
-3		76 78	CSE2
		89	
		41	
			---
		1 42	
			(1)
-4			Below
Measr	-Raters	+Performances	CSEW_A

Table 68 shows the Fair Average and CSE level estimation for all 120 Aptis Performances. The table can be interpreted in the same way as for Aptis Speaking.

Table 68: Fair average and level estimation for Aptis Writing performances

Order	Count	Fair Avg	Level	Aptis Score	Order	Count	Fair Avg	Level	Aptis Score
1	14	1.42	1	12	**61**	14	14.62	15	48
2	14	14.97	15	46	**62**	14	13.54	14	46
3	14	3.7	4	22	**63**	14	3.49	3	24
4	14	2.26	2	16	**64**	14	13.25	13	44
5	14	6.43	6	34	**65**	14	11.63	12	44
6	14	7.2	7	28	**66**	14	7.41	7	40
7	14	6.36	6	30	**67**	14	13.18	13	46
8	14	2.4	2	18	**68**	14	6	6	26
9	14	10.17	10	40	**69**	14	12.37	12	42
10	14	10.46	10	42	**70**	14	11.49	11	42
11	14	10.1	10	44	**71**	14	7.98	8	34
12	14	6.43	6	24	**72**	14	8.26	8	26
13	14	8.82	9	42	**73**	14	2.26	2	14
14	14	5.57	6	36	**74**	14	6.78	7	38
15	14	13.33	13	48	**75**	14	2.26	2	12
16	14	7.34	7	26	**76**	14	1.91	2	14
17	14	7.55	8	38	**77**	14	2.33	2	20
18	14	6.92	7	34	**78**	14	1.98	2	14
19	14	5.07	5	34	**79**	14	2.12	2	20
20	14	8.68	9	42	**80**	14	2.05	2	20
21	14	6.57	7	28	**81**	14	2.19	2	18
22	14	3.7	4	22	**82**	14	3.27	3	22
23	14	5.86	6	34	**83**	14	3.78	4	18
24	14	9.74	10	40	**84**	14	4.86	5	24
25	14	9.1	9	40	**85**	14	5.29	5	34
26	14	3.99	4	28	**86**	14	2.76	3	20
27	14	5.14	5	36	**87**	14	2.69	3	20
28	14	5.22	5	36	**88**	14	8.96	9	38
29	14	5.36	5	32	**89**	14	1.77	2	12
30	14	6.85	7	16	**90**	14	2.4	2	14
31	14	5.43	5	32	**91**	14	4.28	4	28
32	14	4.21	4	20	**92**	14	10.61	11	44
33	14	6.36	6	30	**93**	14	12.22	12	44
34	14	2.69	3	14	**94**	14	5	5	26
35	14	3.78	4	24	**95**	14	12	12	44
36	14	7.55	8	38	**96**	14	7.06	7	30
37	14	6.64	7	36	**97**	14	2.33	2	18
38	14	5.36	5	26	**98**	14	3.41	3	32
39	14	9.24	9	30	**99**	14	2.54	3	18
40	14	7.48	7	40	**100**	14	2.69	3	14
41	14	1.56	2	12	**101**	14	5.07	5	26
42	14	1.42	1	12	**102**	14	2.62	3	16
43	14	5.5	6	30	**103**	14	5.07	5	24
44	14	4.5	5	24	**104**	14	5.64	6	28
45	14	4.07	4	22	**105**	14	5.5	6	32
46	14	12.67	13	42	**106**	14	5.79	6	32
47	14	15.04	15	46	**107**	14	5.79	6	32
48	14	14.55	15	46	**108**	14	12.08	12	44
49	14	14.69	15	50	**109**	14	5.29	5	22
50	14	9.17	9	38	**110**	14	2.26	2	14
51	14	4.07	4	18	**111**	14	3.78	4	18
52	14	14.9	15	50	**112**	14	6	6	28
53	14	13.91	14	48	**113**	14	7.77	8	38
54	14	13.47	13	48	**114**	14	4.21	4	12
55	14	11.41	11	46	**115**	14	7.2	7	36
56	14	12.3	12	46	**116**	14	10.03	10	40
57	14	14.48	14	48	**117**	14	7.48	7	36
58	14	14.19	14	48	**118**	14	4.86	5	30
59	14	14.33	14	50	**119**	14	3.41	3	22
60	14	14.05	14	48	**120**	14	4.93	5	16

Table 69 presents the mean and median Aptis scale scores for all performances allocated to each of the overall CSE levels in the table. Note that no performances were allocated to the CSE 7 level, one of the target levels for this linking project for Aptis. Table 70 displays the cutoff estimates for the three methods described for Speaking, starting with the borderline approach. As already noted, as no performances were allocated to the CSE 7 level, it was not possible to estimate a cutoff point for CSE 7 for Aptis Writing.

Table 69: Mean and median Aptis Writing scale scores for each CSE level

Level	Mean	Median
CSE 2	17.5	18
CSE 3	26.9	28
CSE 4	34.4	36
CSE 5	43	44
CSE 6	47.2	48
CSE 7	N/A	N/A
CSE 8	N/A	N/A

Table 70: Cutoff estimates for CSE 3 to CSE 7 for Aptis Writing

Borderline		Collapsed categories	
Level		Mean of mean	Midpoint of median
CSE 3	20.7	22.2	23
CSE 4	31.4	30.65	32
CSE 5	39.6	38.7	40
CSE 6	44.7	45.1	46
CSE 7	N/A	N/a	N/A

7.2.5.2.3 Internal validity.

As with Speaking, inter-rater correlations were calculated using the second round judgements for all raters. For Writing, as all raters rated all performances, the inter-rater correlations were calculated on the full set of 120 performances. The mean inter-rater correlation was 0.81. As with Speaking, the MFRM analysis is a key part of ensuring the robustness of level estimates for the performances used to calculate the cutoff estimates for the AJM procedure.

7.2.5.3 IELTS Writing

7.2.5.3.1 Analysis procedure using MFRM

The same analysis approach as described for Aptis was employed for IELTS. The 16 raters rated all 120 writing performances used in this project. The rating scale employed was slightly modified to reflect the CSE levels targeted for the IELTS test in this project, as with Speaking. Table 71 shows the rating scale with the CSE levels, split into low, mid and high sublevels.

Table 71: Rating scale used for IELTS Writing samples

Rating	Level	Rating	Level	Rating	Level	Rating	Level
1	Below 3	7	5 Low	13	7 Low	19	CSE 9
2	CSE 3	8	5 Mid	14	7 Mid		
3	3 High	9	5 High	15	7 High		
4	4 Low	10	6 Low	16	8 Low		
5	4 Mid	11	6 Mid	17	8 Mid		
6	4 High	12	6 High	18	8 High		

7.2.5.3.2 Results

With IELTS Writing, four analysis runs were required until all remaining raters showed adequate fit. One rater was dropped in each of the first three analysis runs: in the first run R13 (with an infit Mean Square of 1.80), in the second run R11 (with an Infit Mean Square of 1.55), and in the third run, R8 (with an Infit Mean Square of 1.59). Figure 54 shows the rater measurement report from the fourth analysis run in which all remaining 13 raters showed adequate levels of fit. Figure 55 shows the facet map, which can be interpreted in the same way as described for Speaking and Aptis Writing above. Table 72 presents Fair Average estimates for each performance and the final level allocation based on the rounded Fair Average result. The original IELTS band the performance was allocated is also displayed.

Figure 54: Rater measurement report for IELTS Writing

CSE_W IELTS Round2 D4 : delete further misfitting rater 23/09/2018 11:49:41
Table 7.1.3 Raters Measurement Report (arranged by N).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBis	Exact Agree. Obs %	Exact Agree. Exp %	Group	Nu	Raters
1104	120	9.20	9.15	-.14	.05	.78	-1.8	.78	-1.8	1.14	.84	16.4	16.2	1	1	R1
1063	120	8.86	8.81	-.02	.05	1.34	2.4	1.34	2.4	.66	.70	15.1	16.5	1	2	R2
998	120	8.32	8.27	.17	.05	1.17	1.3	1.19	1.4	.82	.79	14.6	16.3	1	3	R3
1029	120	8.57	8.53	.08	.05	.51	-4.6	.51	-4.7	1.48	.86	17.2	16.5	1	4	R4
1192	120	9.93	9.88	-.40	.05	1.05	.4	1.07	.5	.91	.77	14.3	14.6	2	5	R5
1147	120	9.56	9.51	-.27	.05	.80	-1.6	.78	-1.8	1.22	.74	15.5	15.6	2	6	R6
1194	120	9.95	9.90	-.41	.05	1.36	2.5	1.36	2.5	.67	.72	12.7	14.6	2	7	R7
1018	120	8.48	8.44	.11	.05	.92	-.5	.92	-.5	1.07	.71	16.9	16.4	3	9	R9
973	120	8.11	8.07	.25	.05	1.50	3.4	1.50	3.3	.53	.69	15.5	15.9	3	10	R10
1086	120	9.05	9.00	-.09	.05	1.16	1.2	1.14	1.0	.92	.81	16.4	16.4	3	12	R12
950	120	7.92	7.88	.32	.06	1.34	2.4	1.34	2.4	.68	.84	14.3	15.5	4	14	R14
974	120	8.12	8.08	.24	.05	.48	-4.9	.49	-4.9	1.54	.89	17.4	16.0	4	15	R15
994	120	8.28	8.24	.18	.05	.44	-5.5	.43	-5.7	1.54	.86	17.8	16.2	4	16	R16
1055.5	120.0	8.80	8.75	.00	.05	.99	-.4	.99	-.4		.79				Mean (Count: 13)	
80.0	.0	.67	.66	.24	.00	.35	3.0	.35	3.0		.07				S.D. (Population)	
83.3	.0	.69	.69	.25	.00	.36	3.1	.37	3.1		.07				S.D. (Sample)	

Model, Populn: RMSE .05 Adj (True) S.D. .23 Separation 4.23 Strata 5.98 Reliability (not inter-rater) .95
Model, Sample: RMSE .05 Adj (True) S.D. .24 Separation 4.41 Strata 6.22 Reliability (not inter-rater) .95
Model, Fixed (all same) chi-square: 246.6 d.f.: 12 significance (probability): .00
Model, Random (normal) chi-square: 11.4 d.f.: 11 significance (probability): .41
Inter-Rater agreement opportunities: 9360 Exact agreements: 1469 = 15.7% Expected: 1489.3 = 15.9%

Figure 55: Facet map for IELTS Writing

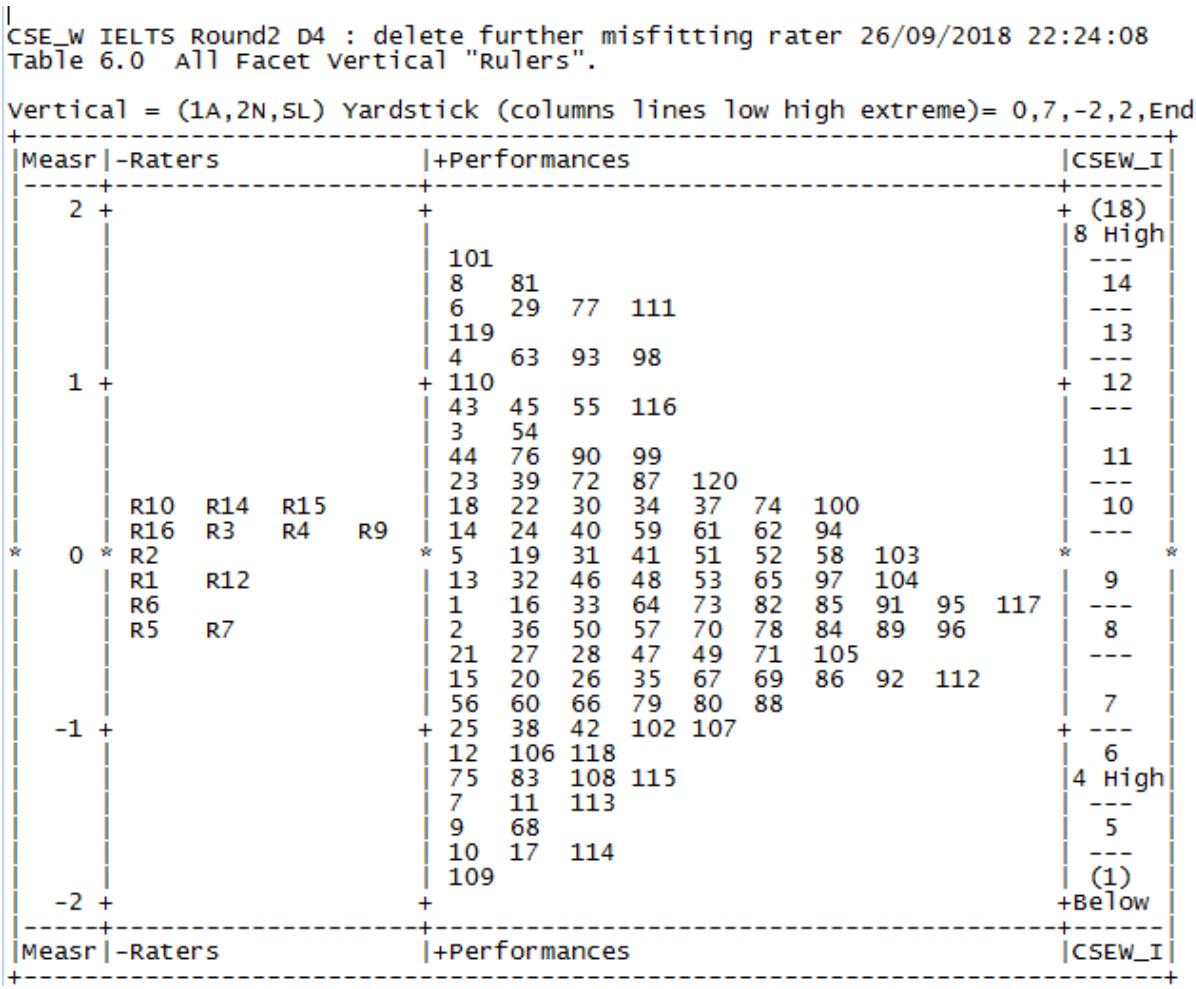


Table 72: Fair average and level estimation for IELTS Writing performances

Order	Count	Fair Avg	Level	IELTS Band	Order	Count	Fair Avg	Level	IELTS Band
1	13	8.61	9	6	**61**	13	9.46	9	6.5
2	13	8	8	6	**62**	13	9.69	10	6.5
3	13	11.44	11	6.5	**63**	13	12.76	13	8.5
4	13	12.6	13	8	**64**	13	8.61	9	7.5
5	13	9.31	9	7.5	**65**	13	8.69	9	6.5
6	13	13.39	13	7	**66**	13	7.09	7	5.5
7	13	5.46	5	5	**67**	13	7.16	7	6
8	13	13.71	14	8	**68**	13	5.07	5	5
9	13	4.91	5	4.5	**69**	13	7.47	7	5.5
10	13	4.6	5	4.5	**70**	13	8.15	8	6
11	13	5.23	5	4.5	**71**	13	7.62	8	6
12	13	6.16	6	5.5	**72**	13	10.38	10	7
13	13	9	9	6	**73**	13	8.38	8	5.5
14	13	9.69	10	6	**74**	13	9.92	10	6
15	13	7.39	7	5.5	**75**	13	5.7	6	5.5
16	13	8.53	9	6.5	**76**	13	10.83	11	7
17	13	4.52	5	5	**77**	13	13.47	13	6
18	13	9.92	10	6.5	**78**	13	8.23	8	6
19	13	9.38	9	6	**79**	13	6.86	7	5.5
20	13	7.16	7	5.5	**80**	13	7.09	7	5.5
21	13	7.85	8	5.5	**81**	13	14.1	14	8.5
22	13	10.08	10	6.5	**82**	13	8.3	8	6
23	13	10.3	10	6.5	**83**	13	5.77	6	5.5
24	13	9.46	9	6.5	**84**	13	8	8	5.5
25	13	6.71	7	5	**85**	13	8.38	8	5.5
26	13	7.31	7	5.5	**86**	13	7.24	7	5.5
27	13	7.62	8	5.5	**87**	13	10.3	10	6.5
28	13	7.54	8	6	**88**	13	6.93	7	5.5
29	13	13.39	13	7	**89**	13	8.15	8	6
30	13	10.15	10	6	**90**	13	10.99	11	7
31	13	9.15	9	5.5	**91**	13	8.38	8	6.5
32	13	8.69	9	6	**92**	13	7.47	7	5.5
33	13	8.3	8	6	**93**	13	12.76	13	8.5
34	13	10	10	7	**94**	13	9.62	10	5.5
35	13	7.16	7	5.5	**95**	13	8.46	8	5.5
36	13	7.92	8	6	**96**	13	8	8	5.5
37	13	9.92	10	5.5	**97**	13	8.69	9	6
38	13	6.47	6	5.5	**98**	13	12.68	13	7.5
39	13	10.53	11	7	**99**	13	10.83	11	7
40	13	9.85	10	5.5	**100**	13	10.08	10	6.5
41	13	9.38	9	7	**101**	13	14.18	14	8.5
42	13	6.71	7	5.5	**102**	13	6.55	7	4.5
43	13	11.82	12	7	**103**	13	9.31	9	5
44	13	11.06	11	6.5	**104**	13	8.76	9	5
45	13	11.59	12	7	**105**	13	7.62	8	5
46	13	8.76	9	6	**106**	13	6.09	6	4.5
47	13	7.85	8	6.5	**107**	13	6.55	7	5
48	13	8.69	9	6	**108**	13	5.54	6	4.5
49	13	7.69	8	5.5	**109**	13	4.37	4	4.5
50	13	8.15	8	6	**110**	13	11.98	12	8
51	13	9.38	9	6	**111**	13	13.47	13	7
52	13	9.15	9	6.5	**112**	13	7.39	7	4.5
53	13	8.92	9	6.5	**113**	13	5.3	5	4.5
54	13	11.21	11	7	**114**	13	4.6	5	5
55	13	11.67	12	7	**115**	13	5.77	6	5
56	13	7.01	7	5.5	**116**	13	11.82	12	8
57	13	7.92	8	5	**117**	13	8.46	8	5
58	13	9.23	9	5.5	**118**	13	6.01	6	4.5
59	13	9.85	10	6.5	**119**	13	12.84	13	8.5
60	13	7.09	7	5.5	**120**	13	10.46	10	7

The mean and median IELTS band scores for performances allocated to each level are shown in Table 73. Note that no performances were allocated to CSE levels 3 or 4, or for CSE 8 or 9. This means that cutoff estimates were not able to be produced for CSE 4 and CSE 8, two levels which were originally targeted for IELTS. Cutoff estimates were set for CSE 5, CSE 6, and CSE 7 using all three methods described above for Aptis and IELTS Speaking and Aptis Writing, and these are shown in Table 74.

Table 73: Mean and median IELTS scale scores for each CSE level

Level	Mean	Median
CSE 3	N/A	N/A
CSE 4	N/A	N/A
CSE 5	4.9	5
CSE 6	5.8	5.5
CSE 7	6.6	6.5
CSE 8	N/A	N/A
CSE 9	N/A	N/A

Table 74: Cutoff estimates for CSE 3 to CSE 7 for Aptis Writing

Level	Borderline	Collapsed categories	
		Mean of mean	Midpoint of median
CSE 4	N/A	N/A	N/A
CSE 5	5.3	5.35	5.25
CSE 6	6.2	6.2	6
CSE 7	7.5	7.2	7.25
CSE 8	N/A	N/A	N/A

7.2.4.3.3 Internal validity

Inter-rater correlations were calculated using the second round judgements for all raters. For IELTS Writing, as all raters rated all performances, the inter-rater correlations were calculated on the full set of 120 performances. The mean inter-rater correlation was 0.60. As already noted for Speaking and Writing, the MFRM analysis is a key part of ensuring the robustness of level estimates for the performances used to calculate the cutoff estimates for the AJM procedure. This procedure has allowed us to maximise the multiple-rater design to elicit ratings on all 120 performances in a fully crossed, robustly linked design. MFRM has ensured that only ratings from raters demonstrating adequate fit are included in the final analysis. For IELTS Writing, although three raters were dropped due to misfit, this still left a pool of 13 raters, all of whose ratings could be used in estimating final measures for the performance. It is worth noting again that although we applied the Infit Mean Square threshold consistently, all of the three raters dropped had Infit results between 1.5 and 2.0, which as noted above, a range which as noted previously has been suggested is not degrading for measurement.

7.2.6 Procedural validity

As described in Section 5.2.5, one source of procedural validity evidence is feedback obtained from participants immediately following the relevant standard setting sessions. Accordingly, three questionnaires were completed by the panellists: one after the Listening, one after the Reading and one after the Speaking and Writing sessions. A single questionnaire was used after the Speaking and Writing sessions since these were carried out together over the same period and using the same methodology.

During all standard-setting sessions, participants were instructed that they would be provided with the opportunity to reflect on the process and provide feedback at the end of the procedures, and that they should give their opinions at this time rather than during standard setting itself. The aim of this was to maintain the flow of activities and limit comments and questions to clarification and immediate implementation of the procedures.

The main areas covered in the questionnaire feedback are the overall composition of the expert panels, the extent to which participants were satisfied that they were able to follow the documented procedure, the sufficiency of the training provided, and the opportunity for discussion. For the purposes of conciseness, the questionnaire results are not dealt with in detail but are outlined holistically below. A detailed summary of results is provided Appendix D. Thus, the results below provide an overall picture of panellists' ability and readiness to give reasonable judgements within the context of the standard-setting procedure.

The expert panellists were selected to represent a population of relevant professionals within the Chinese context. The majority of panellists were female and the first language of the vast majority (81%) was Chinese. As shown in Figure 58, each standard-setting group had representation from people with teaching experience in the Elementary School, High School, University and Business categories. Most of the participants had experience at university level. The second largest category is Other, which was found to consist almost entirely of assessment-related roles, including examiner, item writer and test researcher. Respondents stated that these roles were often performed concurrently with teaching duties at the same institution.

Figure 56: Gender of standard-setting participants

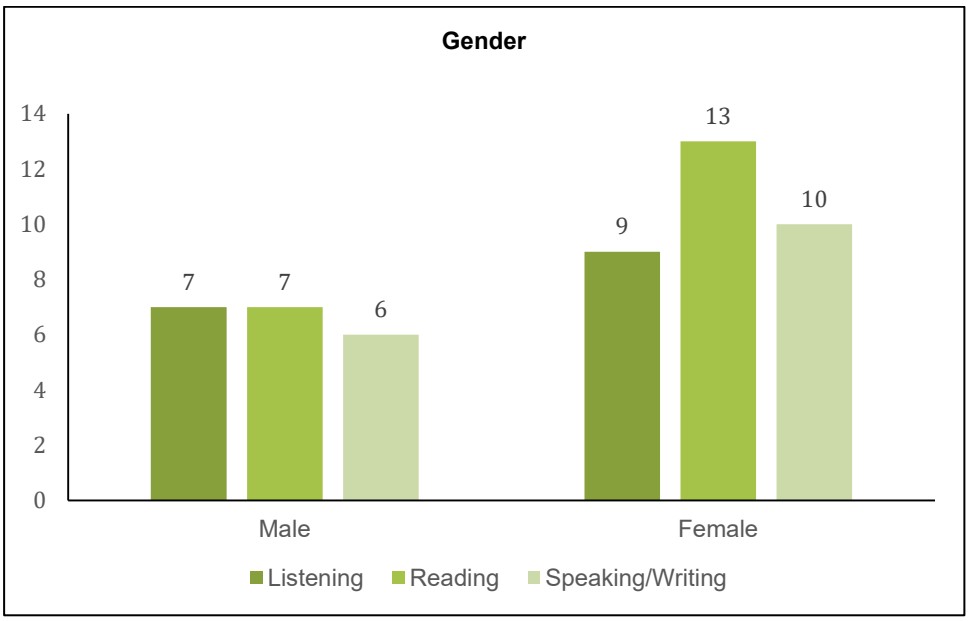


Figure 57: First language of standard-setting participants

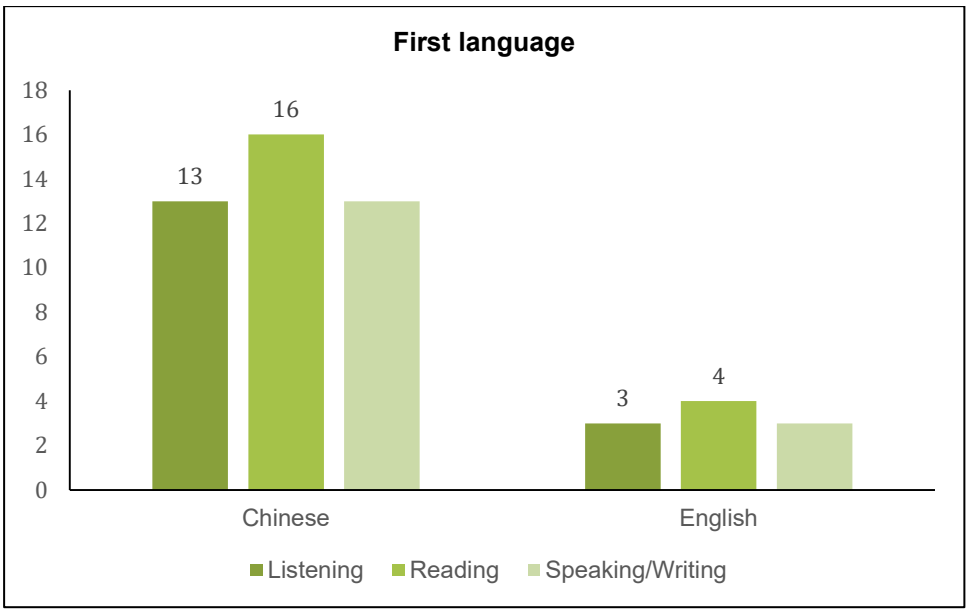
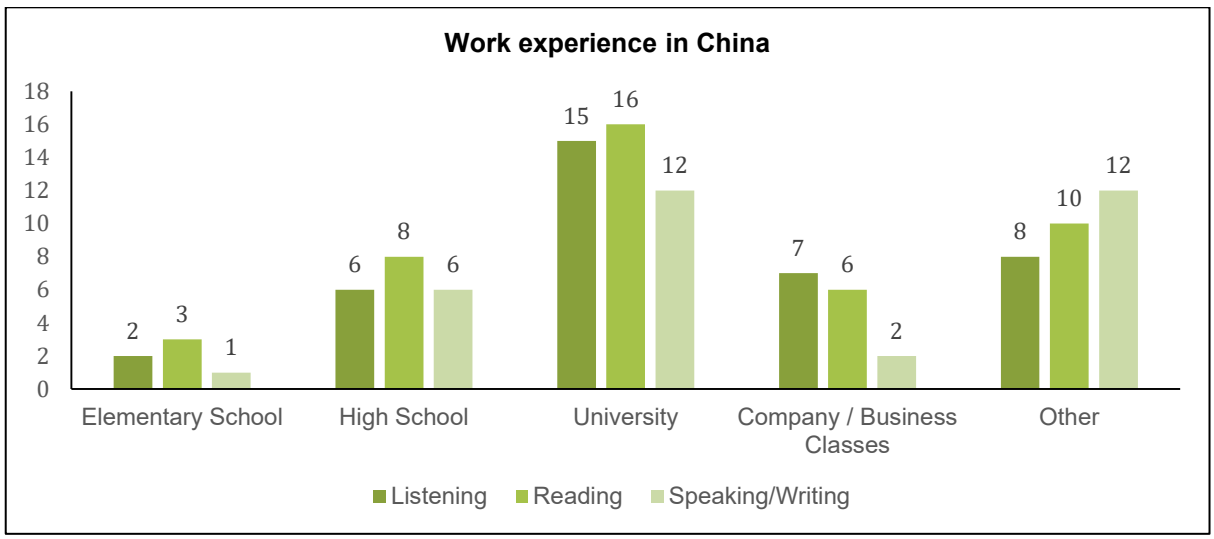


Figure 58: Participant work experience in teaching and assessment in China



The participants began the standard-setting processes with various degrees of knowledge of the CSE. Most participants had a high level of familiarity with the CSE. This number increased after the Listening standard setting, as the CSE was released publicly after this time and many of the participants in Listening were also panellists for the other components, meaning that they came to the later components with experience of the Listening process.

As described above, before standard setting for each component, participants were provided with a self-study preparation booklet containing detailed descriptions of the CSE and familiarisation activities in order to facilitate the training process. The results in Figures 59 and 60 show that the preparation materials had a beneficial effect on participants' understanding of important concepts in advance of standard setting, with 94% across components agreeing that they had a clear understanding of the purpose of the project and the structure of the CSE. Only three people disagreed in each case, for Listening, which it is to be noted was the initial pilot study in this process.

Verbal feedback from participants indicated that, in some cases, more clarity was required regarding the role of the preparation booklet in the procedure, i.e., as to whether it was compulsory or served

merely as supplementary material. Expectations over required completion of the subsequent Reading and Speaking/Writing booklets were emphasised as a result. Overall, results provide strong evidence that the preparation materials were successful and contributed to procedural validity.

Figure 59: Effect of the preparation booklet on participant understanding of the project purpose

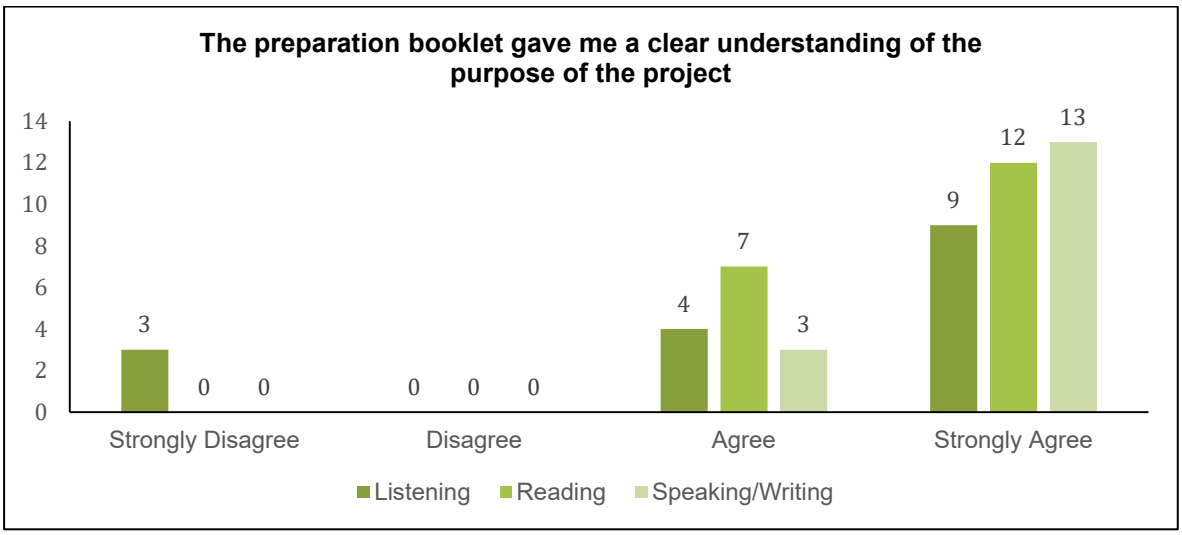
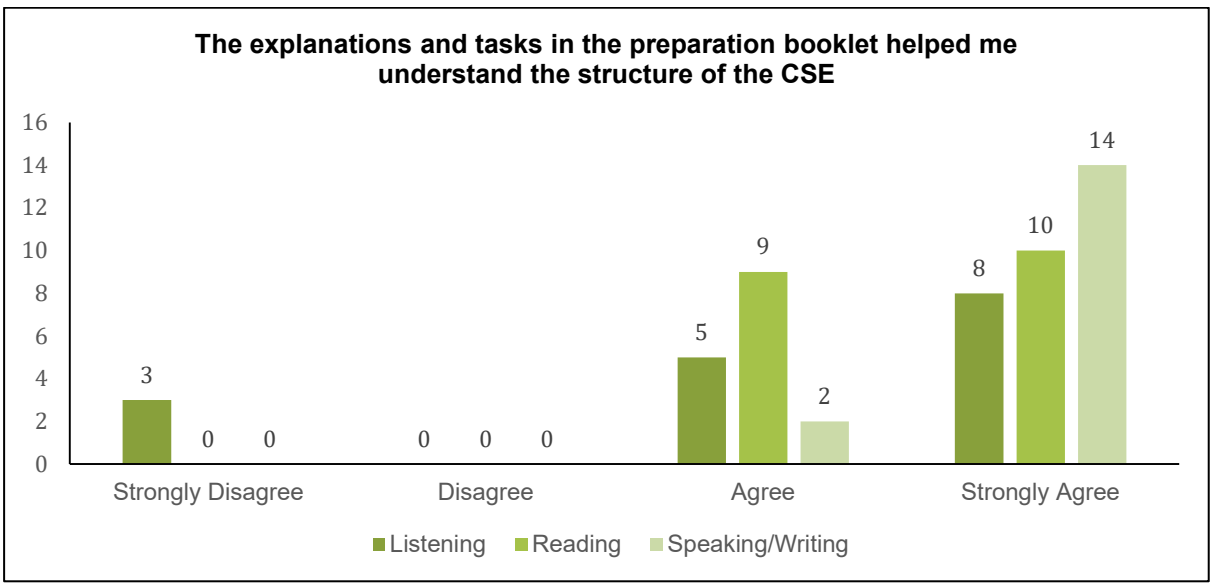


Figure 60: Effect of the preparation booklet on participant understanding of the CSE



In the face-to-face training sessions, participants were provided with an explanation of the different methodologies to be used and Figures 61 and 62 display the perceived readiness of participants to make judgements during the Listening and Reading standard setting using the Basket and Angoff Methods, respectively. In the case of both methods, 91% of respondents indicated that the explanations of these methodologies had prepared them adequately for the tasks. Only three people were in disagreement in each case, suggesting that overall participants felt confident in using the methodology to make their judgements. An even more positive picture emerges from the Speaking and Writing standard setting, in which the Analytical Judgement Method was used. All respondents agreed that the explanation of the method enabled them to carry out the rating task.

Figure 61: Participant understanding of the Basket Method

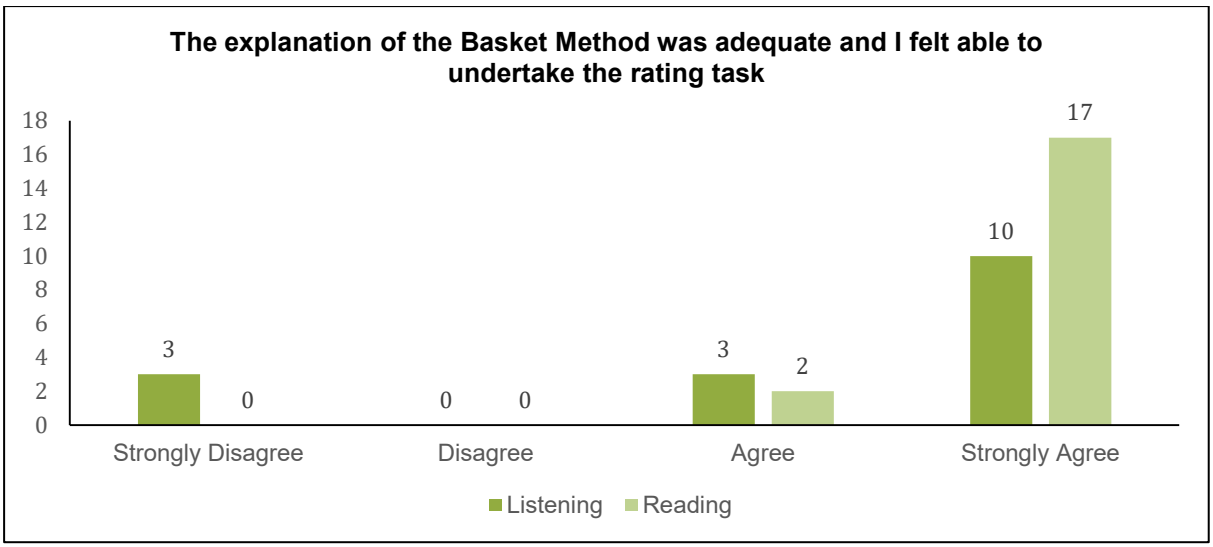


Figure 62: Participant understanding of the Angoff Method

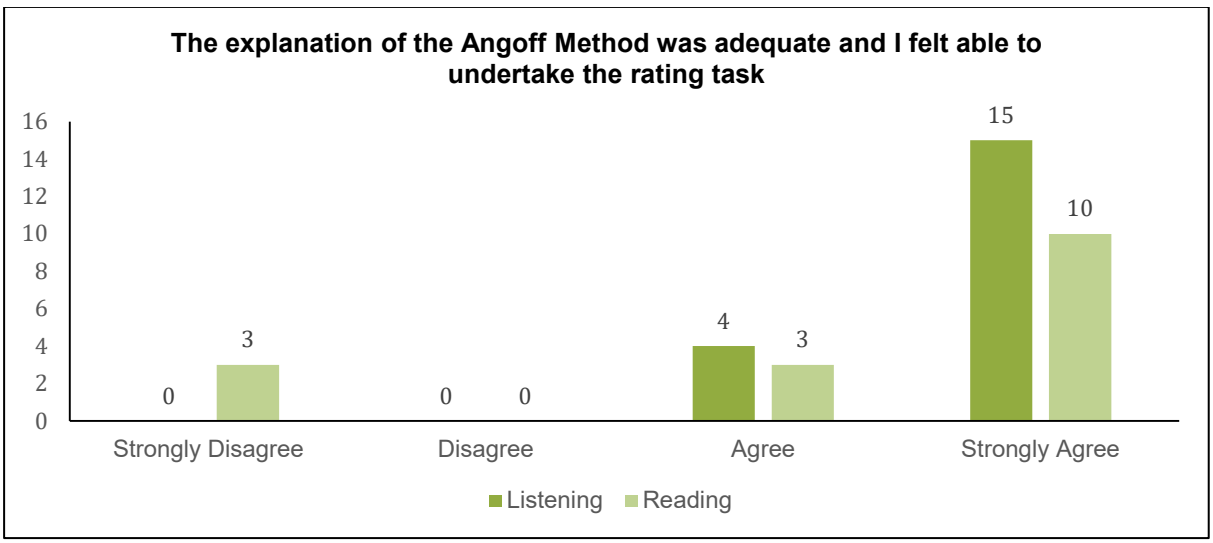
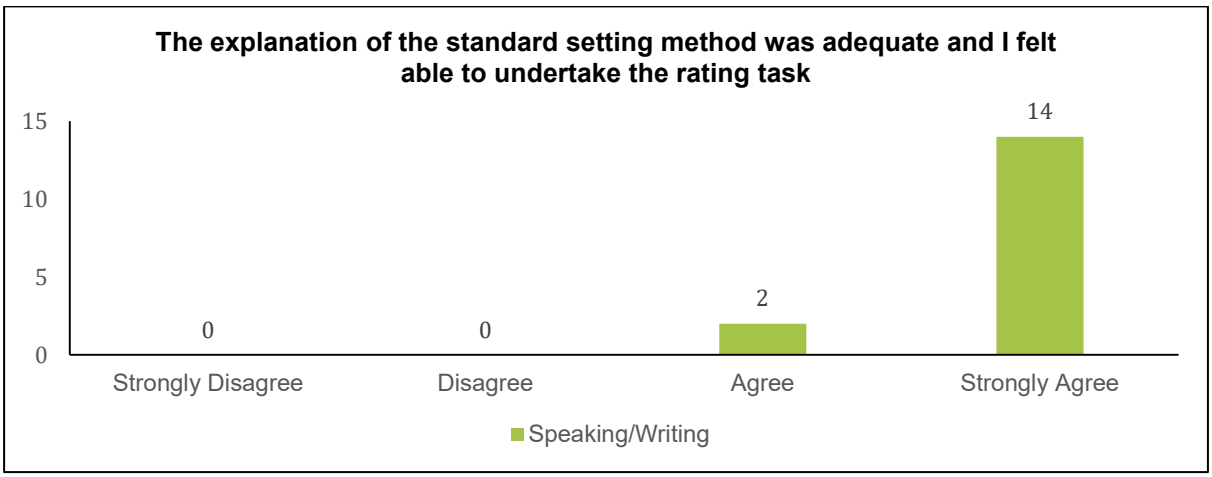


Figure 63: Participant understanding of the Analytical Judgement Method



Participants were asked about the discussions which took place as part of the standard setting procedures. Figure 64 shows that a very large majority (94%) felt that they had benefitted from

discussion of the CSE during the familiarisation and standardisation stages of the process. The same proportion also agreed that they had the opportunity to contribute to discussion (Figure 65). This would suggest that the amount of discussion during the sessions was sufficient and that discussion was also productive.

Figure 64: Participant discussion of the CSE

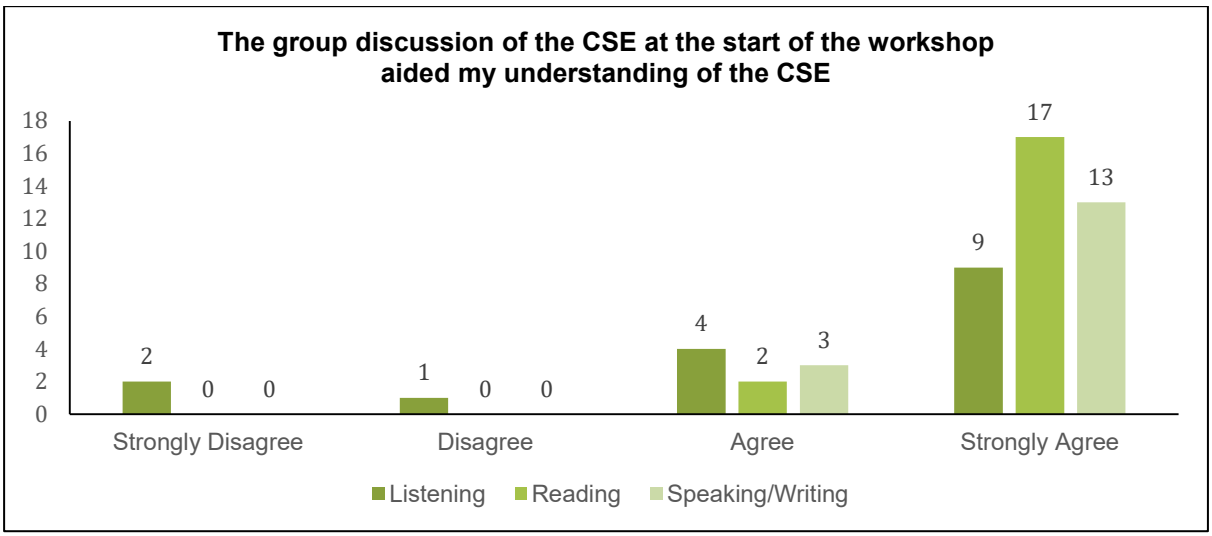


Figure 65: Opportunities for participant discussion

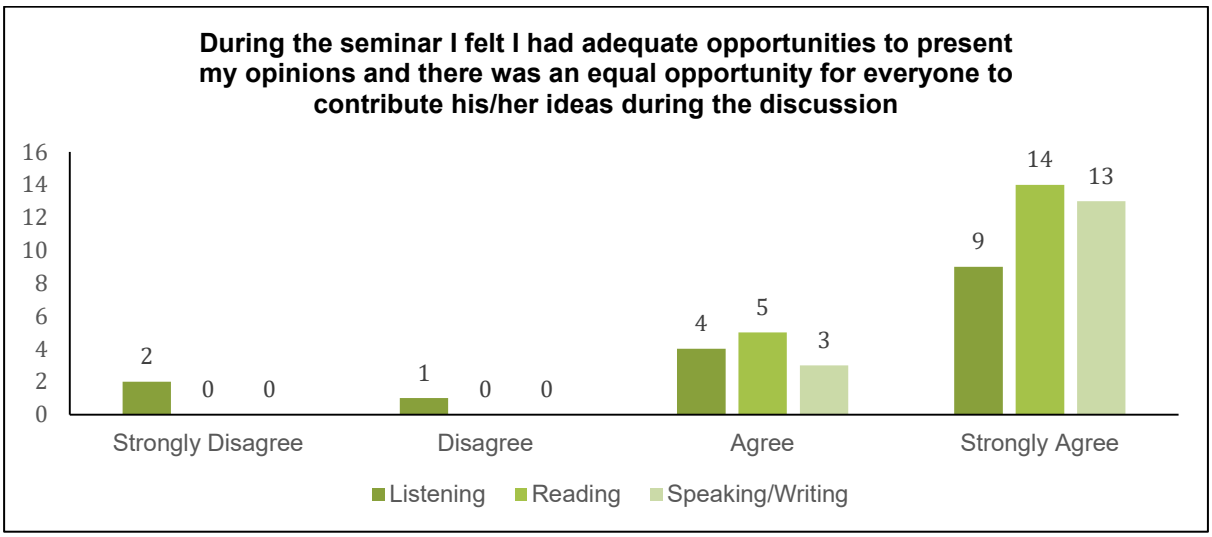
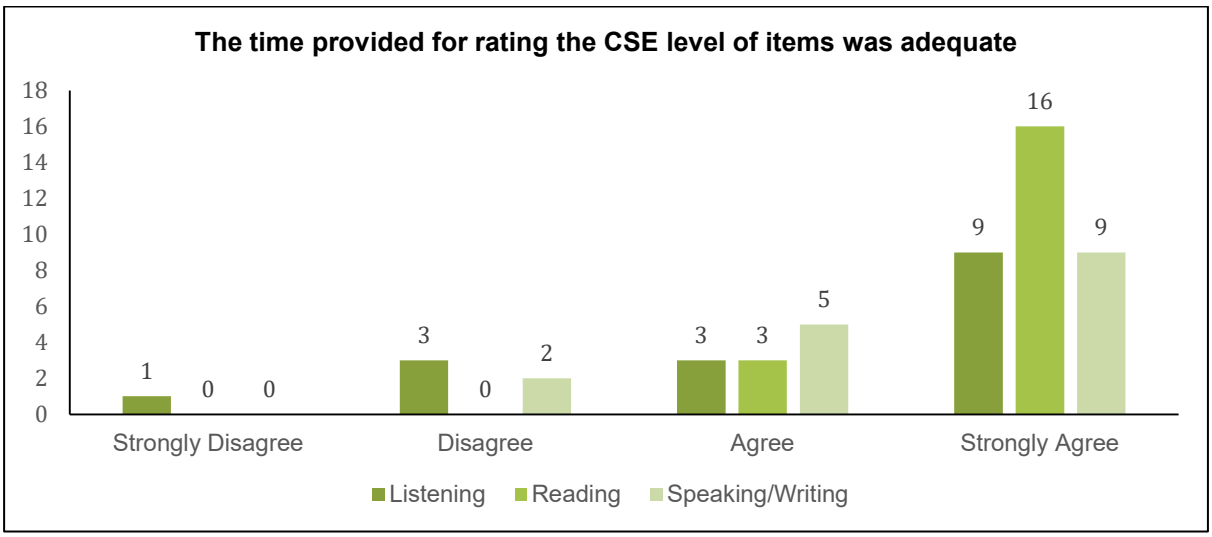


Figure 66 shows how participants felt about the amount of time they were given to complete the rating tasks. A very high proportion (88%) agreed that it was adequate, with all participants in agreement for Reading. Only four disagreed for Listening and two disagreed for Speaking/Writing. This suggests that overall, the time provided was sufficient although it should be noted that the Listening and Speaking/ Writing rating tasks appeared to place more demands on raters in this respect.

Figure 66: Time provided for rating tasks



To conclude, the questionnaire responses are a strong indication that the members of the three expert panels felt adequately prepared to give CSE ratings and had the opportunity to further clarify their understanding in terms of the CSE scales, relevant English tests and methodologies used in this study. As such, they provide evidence of the procedural validity of the standard setting carried out in this study.

7.3 External validation

7.3.1 Overview

As described in the methodology section, external validation was built in to the methodology from the beginning. As also noted, from the outset, panel-based standard setting was placed at the centre of the data collection for setting cutoffs. To supplement this approach, an examinee-centred standard setting approach which would utilise test-taker data in conjunction with teacher judgements was built into the internal validation stage. However, it was recognised from the outset that the collection of data through this method would face severe limitations because the CSE would only be rolled out officially after the data collection, and so teachers and students would not be familiar with the scales or experienced in its use. As such, this part of the project was seen as an important pilot opportunity and as also noted in the methodology section, would potentially provide important insights into the perceptions of important stakeholders, such as teachers.

The following sections provide an overview of how such an external validation project can be approached and the preparations that needs to be involved in ensuring adequate training and familiarisation for participants. External validation, particularly the collection of teacher judgements and student test data, should be seen as a longer-term, ongoing research project, particularly in the case of such a new set of standards such as the CSE, and changes in perceptions and familiarity would be expected as implementation proceeds.

The pilot phase of this external validation approach has provided the project with direct interaction with a large number of teachers and students. Due to the anticipated issues with lack of familiarisation in actual implementation, the quantitative data collected from this stage of the project was not used in the final estimation of cutoffs in the recommendations section, and as such is not presented in this report. A description of the data collection procedures is included as a useful guide for how this kind of procedure can be operationalised. As noted in the recommendations section, this data will provide a potentially rich source of follow-up studies, including impact studies to investigate the perceptions of the CSE of participants who took part.

A separate strand of external validation is the triangulation of alignment studies carried out for CSE and the individual examinations in this study with another external, international proficiency

framework, the CEFR. Preliminary results for this strand of external validation are reported in Section 7.3.1.2.

7.3.2 Test data and teacher judgements through examinee-centred standard setting

7.3.2.1 Sampling plan

7.3.1.1.1 Sampling principles

First, the student samples, to some extent, should be representative of the test population of the exam programs concerned. The main features of the test population, such as geographical distribution, educational background, proficiency levels, should be considered in selecting the samples.

Second, the sample size should meet the minimum requirement of statistical analysis. Therefore, a sample size of 300 examinees was targeted for each test in this research.

7.3.2.1.2 Sampling procedures

Step1: Region selection. China has 34 provincial regions, including 23 provinces, five autonomous regions, four municipalities and two special administrative regions. However, it was almost impossible to cover all the provinces or regions, so sample provincial regions were selected according to the geographical division of China as East China, South China, North China and West China. For each area, one or two provinces were chosen for the sake of convenience.

Step 2: School selection. In order to cover a wide range of IELTS and Aptis test population, different categories of schools were selected, e.g. public high schools, private high schools, universities, and language training schools. For the IELTS sample, 80% focused on university students while 20% focused on high school students. For the Aptis sample, 80% were high school students while 20% were university students. Different types of university were also covered to ensure a wider distribution of students' proficiency levels.

Step 3: Student selection. There were three criteria for choosing students. First, they needed to take the live tests (either IELTS or Aptis) rather than mock tests. Second, they would be willing to participate in the research. Third, they needed to take the test in any session between February 2018 and June 2018.

Step 4: Teacher selection. The teacher participants all came from the schools or universities from which the students were chosen as participants. To reduce the workload, as well as ensure the quality of judgement, it was requested that each teacher evaluate no more than 15 students. All the teachers had to participate in the training workshop onsite or online before making their judgement about the students. The gap between the students taking the test and teacher judgements was a maximum of two months.

7.3.2.2 Training teachers

7.3.2.2.1 Purpose

Training sessions were organised to ensure that teachers could make informed judgements about their students' CSE levels in the areas of listening, reading, writing, and speaking skills. The training was therefore designed to familiarise the teachers with the overall structure of the CSE levels and descriptors, to provide further explanation about salient features differentiating adjacent CSE levels, and to provide participants with practice tasks and examples to illustrate how holistic judgements could be made based on daily observation and how reasonable inferences could be made about their students' abilities in English.

7.3.2.2.2 Training content

Training consisted of three parts: a general introduction to the CSE and its descriptors; further explanation of the selected core descriptors from the overall scales and sub-scales across listening, reading, speaking and writing skill areas; and a series of practice tasks.

In each training session, trainers first provided a general introduction as an overview of the development of the CSE and its major functions in aligning English teaching, learning and assessment in China. This also included a brief introduction to the background and objectives of the linking project. The overall structure of the CSE, including its levels, components and underlying framework of language proficiency, as well as the structure of individual descriptors, were then introduced and elaborated upon to provide a more fine-grained picture of how the CSE can be used to describe English learners at different levels.

At the training sessions, a copy of core descriptors selected by the CSE working group members from the listening, reading, speaking and writing skill areas across all nine CSE levels were provided (see Appendix A). About three to five descriptors were selected for each level with more at the target levels of the Aptis and IELTS tests (CSE3–CSE7). Descriptors most representative of each level closely related to the Aptis and IELTS test tasks were selected from both the overall scales and the sub-scales. In order to facilitate the participants' understanding of the CSE, lists of the salient features of descriptors at different levels in different skill areas were also provided. For comprehension skills, the focus was on the features of input language and the cognitive processing. For production skills, the communicative goals and features of output language were emphasised. By highlighting distinguishing features of the CSE adjacent levels, the aim was to familiarise teachers with the CSE descriptors and help them build images of learners at different CSE levels.

Practice tasks were also designed so that teachers would better understand the descriptors and be able to differentiate CSE levels. Various task types were included, namely ordering descriptors according to their proficiency levels, highlighting salient features in selected descriptors, and evaluating sample student profiles.

7.3.2.2.3 Implementation

Four sessions of on-site training were delivered in Nanjing, Beijing, Shanghai and Guangzhou in March and April 2018. Around 56 teachers attended the on-site training. Those who were not available for the on-site training were advised to attend online self-training in June 2018 before they made their judgements.

7.3.2.2.4 On-site training and data collection

Each on-site training session lasted about three hours and generally followed four steps.

Step 1: Collection of background information. Teachers were required to fill in questionnaires about their background, including personal background information, and their familiarity with the CSE, IELTS and Aptis.

Step 2: General introduction. The trainer gave a general introduction to the CSE in terms of CSE levels, CSE language proficiency framework and the structure of CSE descriptors, as well as a brief overview of the linking project and its objectives.

Step 3: CSE familiarisation. The trainer led the participants through the four target skill areas one by one. For each skill area, teachers were asked first to go through the selected core descriptors across all nine CSE levels. The trainer then focused on the salient features to further explain and clarify how descriptors ascend in proficiency level, followed by the practice tasks which helped teachers better differentiate the core features distinguishing adjacent levels. The tasks were conducted in an interactive manner. Teachers were permitted to ask questions and ask for clarification whenever they felt uncertain about their interpretation of the descriptors. This enabled the trainer to provide quick feedback and further explanation to dispel confusion about the CSE descriptors.

Step 4: Discussion and evaluation. After familiarisation, teachers were first required to evaluate two students they were teaching at the time. It should be noted that those two students may or may not have been on the student list. The teachers were encouraged to talk with other teachers present, and discuss why they had put these students at certain CSE levels in a particular skill area. They were also encouraged to explain any difficulties they had in the judgement process. Through discussion, they were able to raise some common concerns and questions. The trainer then referred back to the descriptors and salient features with the aim of clarifying the guidelines and instructions on the Students Rating Sheet. Since teachers could potentially struggle to balance different descriptors at the same level given their incomplete knowledge of students' performance and potentially jagged proficiency profiles, it was vital to emphasise the importance of making holistic judgements based on the students' observable or potential performance as described in the descriptors at a certain level. Teachers were then given about 15 to 20 minutes to rate about 10 to 15 students in their classes in each skill area.

7.3.2.2.5 Online training and data collection

Due to the logistic constraints in organising all the teachers to attend the on-site training workshops, online training was conducted as an alternative for those who could not participate on site. An online data training and data collection platform was developed for this purpose.

The platform facilitated training in three ways. First, the CSE familiarisation course and exercises could be displayed at the teachers' own pace. Second, teachers' judgement on their students' English language proficiency levels were collected through an electronic version of the onsite questionnaire, which greatly facilitated data collection. Third, relevant background information about the teachers and their students were collected for further research analysis.

The online platform is located on the official website of National Foreign Language Assessment Framework (http://cse.neea.edu.cn). Teachers who attended the online training followed the steps detailed below.

Step 1: Registration. After a quick familiarisation with the linking project background, teachers used their mobile phone numbers to register, so as to facilitate follow-up contact if problematic data emerged. Background information was collected from the teachers, including personal background information, together with their level of familiarity with the CSE, IELTS and Aptis.

Step 2: CSE familiarisation training. Teachers needed to watch a self-training video first, in which the key concepts of the CSE were introduced in terms of the CSE levels, the CSE language proficiency framework, the structure of the CSE descriptors, as well as the salient features of the CSE listening, speaking, reading and writing sub-scales. After watching the training videos, they were required to complete two practice tasks. One task was to rank the order of the CSE writing descriptors. The other was to make judgements about CSE listening levels.

Step 3: Judgements about student proficiency levels. First, teachers were asked to provide background information about the students they were to evaluate. Then, teachers made judgements about those students' CSE levels in listening, reading, writing, and speaking skills. Examples of key features of performance at each CSE level for each of the four main skills were provided to help with the judgement. During the rating process, teachers were allowed to go back to the training session, or make changes to their previous judgements.

7.3.2.2.6 Data collecting outcome

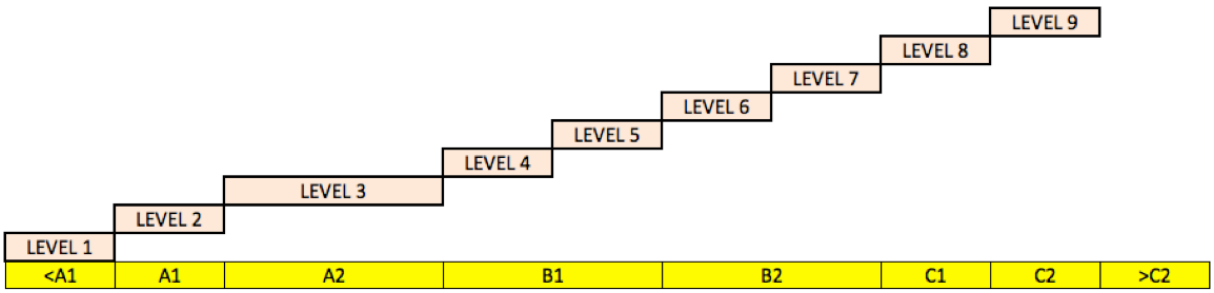
Through the on-site and online data collection, 499 IELTS test-takers' samples with 53 teachers' judgements, and 479 Aptis test-takers' samples with 42 teachers' judgements were successfully collected. These exceeded the original target of 300 samples for each test.

7.3.3 Triangulating claims of relevance to the CEFR

As an additional source of evidence for the external validation stage of the methodology, the relationship between each of the tests in the project and the CEFR, and the relationship between the CSE and the CEFR were examined. The purpose of this evaluation was to triangulate the relationships across the tests with an additional international framework, the CEFR, in order to add additional evidence to support the claim that the linking results from the standard-setting panels are coherent, plausible and defensible. This process drew on several sources of evidence. As a part of the CSE development project, research was carried out to investigate the relationship between CSE levels and CEFR levels (Liu, 2018). This included embedding CEFR descriptors in the data collection carried out to scale the CSE descriptors, then comparing the CSE levels that the descriptors were placed at in relation to their original CEFR levels. In addition, the developers of each of the tests in the project have published information on their alignment with the CEFR, along with supporting information on the rationale and evidence to support these claims. Finally the project team were able to look at the putative cutoff points for CSE levels on each of the test's score scales that had been suggested by the standard-setting panels. The standard-setting results presented for each skill in Section 7.2 include the approximate CEFR level that those cutoff points would fall in according to the developers' original CEFR linking with the tests.

The link between the CSE and the CEFR is appropriately expressed in the Figure 67. Here, the CEFR is shown along the bottom of the scale (note the unequal sizes of the levels – this reflects the reality of the scaling in the CEFR), while the CSE levels are indicated in the top section. It should be remembered that this is somewhat of idealised relationship, representing the *best fit* of CEFR levels in relation to CSE levels based on the various sources of evidence described above. In practice, given that the CSE and CEFR are separate systems with overlapping but nonetheless different perspectives, contexts of use, and development procedures, we would not expect an exact alignment of levels. Indeed the same can be said for alignment results between separate exams and an external standard. Nonetheless, the evidence should be convergent, and support the main alignment claims. The evidence in this case, when comparing the standard-setting results tables in Section 7.2, and the original does indeed provide support for the overall trends in terms of increasing proficiency across the levels of both sets of scales.

Figure 67: Comparing CSE and CEFR levels



7.4 Alignment recommendations for Aptis and IELTS with the CSE

The recommendations for cutoff points on the Aptis and IELTS score scales are presented below as Recommendation 1. These have been derived primarily from reference to the standard-setting results, which the methodology section made clear, was always intended as the central block of evidence for this process. Results from the construct definition phase and the external validation of standard setting have also been presented above to demonstrate a pattern of convergent evidence supporting the alignment in Recommendation 1. Given the comprehensive nature of this project, and the intention for it to inform an ongoing research agenda to support the appropriate implementation of the CSE and the alignment of exams to the CSE, a series of further recommendations have also been listed.

IELTS and Aptis Related Recommendations

The outcomes of the standard-setting panels, in conjunction with other sources of evidence collected as a part of the linking project, have been shown to offer an accurate and consistent estimate of the cutscores relating China's Standards of English Language Ability (CSE) to the IELTS and Aptis tests.

Recommendation 1

The cutscores suggested by the Working Group based on the standard-setting panel results should be adopted with immediate effect. These cutscores may be updated based on the rollout of the CSE and of further planned research into the link between IELTS/Aptis and the CSE.

IELTS	CSE 4	CSE 5	CSE 6	CSE 7	CSE 8
Listening	5	6	6.5	7.5	8.5
Reading	4.5	5.5	6	7	7.5
Speaking	5	5.5	6	6.5	7
Writing	4	5	6	7	7.5
Overall*	**4.5**	**5.5**	**6**	**7**	**8**

** IELTS reports a profile and an overall band score which is derived from averaging the band scores on the profile. This table reflects this approach.*

Aptis	CSE 3	CSE 4	CSE 5	CSE 6	CSE 7
Listening	14	21	29	37	43
Reading	16	26	35	42	46
Speaking	21	29	37	43	47
Writing	22	31	39	45	50

** Aptis reports a profile and an overall score. The overall CEFR/CSE level is estimated by first calculating the CEFR/CSE level independently for each of the four skills and then averaging the CEFR/CSE levels. This table reflects this approach.*

The Working Group recognises the need to recommend an approach to estimating the overall CSE level of a candidate, based on performance on the four skills – note that this overall claim can only be made when the candidate has been tested on all four skills. There are a number of possible ways of doing this, basing the overall claim on:

- an average of the scaled scores
- the sum of the scaled scores
- an average of the CSE level for each skill.

The first two of these approaches risk papering over critical weaknesses in one or more skills and may result in overall estimates that do not necessarily reflect the true ability of the candidate across the four skills.

Recommendation 2

The overall linking claim will be estimated by averaging the CSE level obtained across the four skills – with each skill seen as contributing equally. Where an overall score or level is reported, it should always be accompanied by the performance profile on the four skills upon which the overall score has been calculated.

As indicated in the rationale for Recommendation 1, the Working Group believes that the current cutscores represent an accurate estimate of the link between IELTS / Aptis and the CSE. However, additional research should be carried out which will offer evidence from other perspectives to further confirm the results of the first linking study. This research, in addition to the lessons learnt during the continued rollout of the CSE, will offer us a more holistic picture and may lead to a review of the current claims.

Recommendation 3

The British Council and Cambridge English Assessment partnership (through the IELTS Research Group) continues with its data-based research to confirm or suggest changes to the cutscores proposed in this report.

Even with further research to confirm the current linking results, the long-term acceptance of the link between Aptis and the CSE may be impacted by changes to the test or to the CSE itself. For these reasons, it must be recognised that the maintenance of the claimed link can only be supported by systematic and long-term research.

Recommendation 4

The British Council and Cambridge English Assessment partnership (through the IELTS Research Group) should work to develop a medium and long-term research strategy around the CSE to support its continued successful implementation.

While we feel that it is preferable that a four skills profile is reported, we recognise that an overall score will be required by test users. It is necessary to communicate with stakeholders to identify how best to do this as it is imperative that the reporting system, and the rationale behind that system, is clear and transparent to test users.

Recommendation 5

We recommend that the Steering Group oversees the performance of further research involving the students and teachers who participated in the data-based study in order to ensure that their perceptions and expectations are met in terms of score reporting.

Recommendation 6

The final stage is, in this instance, somewhat problematic, as the CSE has not yet been widely applied across the education system. This means that teachers are not fully familiar with its contents and have not internalised the levels and may not be as accurate in their judgements as we might expect. For this reason, this stage is best seen as a more long-term goal, though evidence was collected across a number of CSE levels where it was felt there was some familiarisation with a group of teachers. The teachers and learners who participated in this initial piloting of collecting teacher judgements of their students' CSE level in addition to test performance data from the students will be consulted as a part of the impact research recommended in Recommendation 5.

REFERENCES

Alderson, J. C. (Ed.) (2002). *Common European Framework of Reference for Languages: Learning, teaching, assessment: case studies*. Strasbourg: Council of Europe.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Alderson, J. C., & Banerjee, J. (2002). Language testing and assessment (Part 2). *Language Teaching*, 35(2), 79–113.

Alderson, J. C., Figueras, N., Kuiper, H., & Nold, G. (2006). Analyzing tests of reading and listening in relation to the Common European Framework of Reference: The experience of the Dutch CEFR Construct Project. *Language Assessment Quarterly*, 3(1), 3–30.

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

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: designing and developing useful language tests*. Oxford: Oxford University Press.

Bachman, L., & Palmer, A. (2010). *Language testing in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing*, 30(4), 441–465.

Bax, S., & Weir, C. J. (2012). Investigating learners' cognitive processes during a computer-based CAE reading text. *Research Notes*, 47, 3–14.

Bechger, T., Kujper, H., & Maris, G., (2009). Standard setting in relation to the Common European Framework of Reference for Languages: the case of the State Examination of Dutch as a Second Language. *Language Assessment Quarterly*, 6(2), 126–150.

Bowers, J. J., & Shindoll, R. R. (1989). *A comparison of the Angoff Beuk, and Hofstee methods for setting a passing score*. Iowa City, IA: ACT Research Report Series, No. 89-2.

Brunfaut, T., & McCray, G. (2015). *Looking into test-takers' cognitive processes while completing reading tasks: A mixed-method eye-tracking and stimulated recall study*. ARAGs Research Reports Online, AR-G/2015/001. London: British Council.

Camilli, G., Cizek, G. J., & Lugg, C. A. (2001). Psychometric theory and the validation of performance standards: History and future perspectives. In G. J. Cizek (Ed.), *Setting performance standards* (pp. 445–476). Mahwah, NJ: Lawrence Erlbaum.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 2–27). London: Longman.

Cizek, G. (1993). Reconsidering Standards and Criteria. *Journal of Educational measurement*, 30 (2), 93–106.

Cizek, G. J. (2001). *Setting performance standards*. Mahwah, NJ: Lawrence Erlbaum.

Cizek, G., & Bunch, M. (2007). *Standard Setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks: Sage Publications.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). An NCME Instructional Module on Setting Performance Standards: Contemporary Methods. *Educational Measurement: Issues and Practice, 23*(4), 31–50.

Clauser, B., Harik, P., Margolis, M., McManus, I. Mollon, J., Chis, L., & Williams, S., (2009). An empirical examination of the impact of group discussion and examinee performance information on judgements made in the Angoff standard-setting procedure. *Applied Measurement Quarterly*, 22(1), 1–21.

Cohen, A., Kane, M., & Crooks, T. (1999). A generalized examinee-centered method for setting standards on Achievement tests. *Applied Measurement in Education*, 12(4), 343–366.

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

Council of Europe (2003). *Relating language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment: Manual: Preliminary pilot version.* Strasbourg: Council of Europe.

Council of Europe. (2004). *Reference supplement to the preliminary version of the manual for relating language examinations to the Common European Framework of References for Languages: learning teaching, assessment*. Strasbourg: Language Policy Division.

Council of Europe. (2009). *Relating language examinations to the Common European Framework of References for Languages: Learning teaching, assessment*. Strasbourg: Language Policy Division.

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

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/2016/002. London: British Council.

Figueras, N., North, B.,Takala, S., Verhelst, N., & Van Avermaet, P. (2005). Relating examinations to the Common European Framework: a manual. *Language Testing*, 22 (3), 1–19.

Geranpayeh, A., & Taylor, L. (eds.) (2013). *Examining listening: research and practice in assessing second language listening*. Studies in Language Testing 35. Cambridge: Cambridge University Press.

Green, D., Trimble, C., & Lewis, D. (2003). Interpreting the results of three different standard-setting procedures. *Educational Measurement: Issues and Practice*, 22(1), 22–32.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards* (pp. 89–116). New York: Routledge.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement*, 24(4), 355–366.

Hurtz, G. M., & Auerbach, M. A. (2003). A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgement consensus. *Educational and Psychological Measurement*, 63, 584–601.

Hurtz, Gr., & Hertz, N. (1999). How many raters should be used for establishing cutoff scores with the Angoff method? A generalizability theory study. *Educational and Psychological Measurement*, 59 (6), 885–897.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 485–514). New York: American Council on Education and Macmillan.

Jaeger, R. (1991). Selection of Judges for Standard Setting. *Educational measurement: Issues and Practice*, 10, (2), 3–10.

Kaftandjieva, F. (2004). Standard setting. In *Council of Europe, reference supplement to the pilot version of the manual for relating language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment (CEF)*. Strasbourg: Language Policy Division.

Kaftandjieva, F. (2009). The Basket method: the bread basket or the basket case of standard setting methods? In N. Figueras & J. Noijons (Eds.), *Linking to the CEFR levels: research perspectives* (pp. 103–109). Arnhem: CITO and EALTA.

Kaftandjieva, F. (2010). *Methods for setting cut-off scores in criterion-referenced achievement tests: A comparative analysis of six recent methods with an application to tests of reading*. Arnhem: CITO and EALTA.

Kane, M. (1998). Choosing between examinee-centered and test centered standard-setting methods. *Educational Assessment*, 5(3), 129–145.

Kane, M. (2001). So much remains the same: conception and status of validation in standard setting. In G. Cizek (Ed.) *Setting Performance Standards*. New York: Routledge.

Khalifa, H. & Weir, C. (2009). *Examining reading: Research and practice in assessing second language reading*. Studies in Language Testing, 29. Cambridge: Cambridge University Press.

Knoch, U. & Frost, K. (2016). *Linking the GEPT Writing Sub-test to the Common European Framework of Reference (CEFR)*. Final report to the Language Training and Testing Center, Taiwan. Language Testing Research Centre, University of Melbourne.

Lim, G.S., Geranpayeh, A., Khalifa, H., & Buckendahl, C.W. (2013). Standard setting to an international reference framework: Implications for theory and practice. *International Journal of Testing*, 13, 32–49.

Linacre, J. M. (2014). *Facets computer program for many-facet Rasch measurement, version 3.71.4*. Beaverton, Oregon: Winsteps.com

Livingston, S., & Zieky, M. J. (1989). A comparative study of standard-setting methods. *Applied Measurement in Education*, 2(2), 121–141.

Liu, J. (2015). Some thoughts on developing China common framework for English language proficiency. *China Examinations*, (1), 7–11.

Liu, J. (2018). *Aligning CSE with CEFR*. Keynote speech delivered at the 4th International Conference on Language Testing and Assessment, Beijing.

Liu, J, & Han, B. (2018). Theoretical considerations for developing use-oriented China's Standards of English Language Ability. *Modern Foreign Languages* (1), 78–90.

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

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749.

Morrow, K. (Ed.). (2004). *Insights from the Common European Framework*. Oxford, England: Oxford University Press.

Nakatsuhara, F. (2014) *A Research Report on the Development of the Test of English for Academic Purposes (TEAP) Speaking Test for Japanese University Entrants – Study 1 & Study 2*. Retrieved from http://www.eiken.or.jp/teap/group/report.html

Norcini, J., Lipner, R., Langdon, L., & Strecker, C. (1987). A comparison of three variations on a standard-setting method. *Journal of Educational Measurement*, 24, 56–64.

North, B. (2000). *The development of a common framework scale of language proficiency*. New York: Peter Lang Publishing, Inc.

North, B. (2007). Response by Brian North. In B. North & T. MacNamara (Chairs), *The CEFR in Europe and beyond: challenges and experiences. Symposium conducted at the 4th European Association of Language Testing and Assessment Conference, Sitges*. Retrieved from http://www.ealta.eu.org.

North, B., Martyniuk, W., & Panthier, J. (2010). Introduction: The manual for relating examinations to the Common European Framework of Reference for Languages in the context of the Council of Europe's work on language Education. In W. Martyniuk, (Ed) *Aligning Tests with the CEFR: Reflections on using the Council of Europe's draft Manual* (pp. 1–17). Cambridge: Cambridge University Press.

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing* 15(2), 217–263.

O'Sullivan, B. (2008). *City & Guilds Communicator IESOL Examination (B2) CEFR linking project: Case study*. Retrieved from: http://cdn.cityandguilds.com/ProductDocuments/International_English/General_English/8984/Additional_documents/8984_Case_study_v1.pdf

O'Sullivan, B. (2010). 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. (2011). Language testing. In J. Simpson (Ed.), *Routledge Handbook of Applied Linguistics*. Oxford: Routledge.

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 (2016). Adapting tests to the local context. In C. Saida, Y. Hoshino & J. Dunlea (Eds.). (2016). *British Council New Directions in Language Assessment: JASELE journal special edition*. Tokyo: British Council.

O'Sullivan, B., & Dunlea, J. (2015). *Aptis General Technical Manual version 1.0*. Aptis Technical Report TR/2015/005. London: British Council.

O'Sullivan, B., & Weir, C. (2011). Test development and validation. In B. O'Sullivan (Ed.), *Language testing: theories and practices* (pp. 13–32). Oxford: Palgrave Macmillan.

Papageorgiou, S. (2010). Investigating the decision-making process of standard setting participants. *Language Testing*, 27(2), 261–282.

Pearson Standards and Quality Office. (2014). *Writing descriptors: Guidelines and best practice*. London: Pearson Publishing Ltd.

Plake, B. S., & Hambleton, R. K. (2000). A standard setting method designed for complex performance assessments: Categorical assignments of student work. *Educational Assessment*, 6(3), 197–215.

Plake, B., Impara, J., & Irwin, P. (2000). Consistency of Angoff-based predictions of item performance: evidence of technical quality of results from the Angoff standard-setting method. *Journal of Educational Measurement*, 37(4), 347–355.

Raymond, M. R., & Reid, J. B. (2001). Who made thee judge? Selecting and training participants for standard setting. In G. J Cizek (Ed.), *Setting performance standards* (pp. 119–157). New York: Routledge.

Reckase, M.D. (2006). A conceptual framework for a psychometric theory for standard setting with examples of its use for evaluating the functioning of two standard-setting methods. *Educational Measurement: Issues and Practice*, 25(2), 14–17.

Shaw, S., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*. Studies in Language Testing 26. Cambridge: Cambridge University Press and Cambridge ESOL.

Tannenbaum, R. J., & Wylie, E. C. (2005*). Mapping English language proficiency test scores onto the Common European Framework*. (ETS Research Rep. No. RR-05-18; TOEFL Research Rep. No. RR–80). Princeton, NJ: ETS.

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

Taylor, L. (2014). *A report on the review of test specifications for the reading and listening papers of the Test of English for Academic Purposes (TEAP) for Japanese University Entrants*. Retrieved from http://www.eiken.or.jp/teap/group/report.html

Van Nijlen, D., & Jansenn, R. (2008). Modelling judgements in the Angoff and Contrasting-Groups methods of standard setting. *Journal of Educational Measurement*. 45 (1), 45–63.

Weir, C. J. (2005). Limitations of the Common European Framework of Reference for Languages (CEFR) for developing comparable examinations and tests. *Language Testing*, 22(3), 281–300.

Weir, C. J., (2014). *A research report on the development of the Test of English for Academic Purposes (TEAP) writing test for Japanese university entrants*. Retrieved from http://www.eiken.or.jp/teap/group/report.html

Weir, C. J., Vidakovic, I., & Galaczi, E. (2013). *Measured constructs: a history of the constructs underlying Cambridge English language (ESOL) examinations 1913–2012*. 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*. Studies in Language Testing: Vol. 41. Cambridge: Cambridge University Press.

Wu, R., Yeh, H., Dunlea, J., & Spiby, R. (2016). *Aptis–GEPT comparison study: Looking at two tests from multiple perspectives using the socio-cognitive model*. British Council Validations Series VS/2016/002. London: British Council.

Yang, H. (2015). Some thoughts on developing a national foreign language testing and assessment system in China. *China Examinations*, (1), 12–15.

Zieky, M. (2001). So much has changed: How the setting of cutscores has changed since the 1980s. In G. J Cizek (Ed.), *Setting performance standards* (pp. 19–52). 380 New York: Routledge.

Appendix A: CSE sub-scales by level

A1 CSE Listening scales by levels

CSE level	Scales	Descriptor
CSE 9	Overall listening comprehension	▪ Can understand spoken discourse on all kinds of topics in all forms; extract main ideas and supporting details; comprehend the implied meaning; and make analyses, inferences, and evaluations.
	Understanding oral description	▪ Can understand detailed oral descriptions of characters and situations using sophisticated vocabulary in literary works and evaluate their role in the development of the theme. ▪ Can understand oral descriptions of experiments in complex reports and extract key points and details.
	Understanding oral narration	▪ Can understand long narratives on all kinds of topics containing low-frequency words and comprehend allusions. ▪ Can understand classic dramas or literary works delivered orally, and comprehend their social, cultural, and historical meaning.
	Understanding oral exposition	▪ Can understand introductory descriptions of academic frontiers regardless of speech rate and comprehend the latest development in the field. ▪ Can understand detailed explanations of complex equipment or precision instruments and comprehend the working mechanism. ▪ Can understand live commentaries on sports events and extract specific information.
	Understanding oral instruction	▪ Can extract key points and procedures from complex technical operation instructions regardless of speech rate and accent.
	Understanding oral argumentation	▪ Can understand complex argumentation on current affairs or social issues regardless of speech rate and accent; and evaluate speakers' opinions and stance. ▪ Can understand debates on political, economic or moral issues regardless of speech rate; and evaluate the effectiveness of strategies used by both sides.
	Understanding oral interaction	▪ Can understand conversations containing low-frequency colloquial expressions or jargon regardless of speech rate; and extract main ideas and supporting details.
CSE 8	Overall listening comprehension	▪ Can understand academic discourse (e.g. lectures, operation instructions) related to his/her own field; and comprehend main ideas and supporting details. ▪ Can follow radio, film, and TV programs regardless of speech rate and accent; and understand the implied meaning of a given discourse, as well as its social and cultural connotations.
	Understanding oral description	▪ Can understand complex oral descriptions of artificial landscapes (e.g. gardens, palaces) regardless of speech rate; and summarise their main features. ▪ Can understand oral descriptions of natural phenomena containing technical terms and summarise their causes and processes.
	Understanding oral narration	▪ Can understand television interviews in language with an accent and identify speakers' opinions and attitudes. ▪ Can extract important information from stories containing colloquial expressions regardless of speech rate. ▪ Can understand well-organised poems delivered orally and appreciate the rhythm and the mood.
	Understanding oral exposition	▪ Can understand highly-informative coverage on popular science regardless of speech rate and summarise the main idea. ▪ Can understand detailed explanations of policy documents and extract specific information.
	Understanding oral instruction	▪ Can understand procedures for experiments in his/her own field regardless of speech rate.

	Understanding oral argumentation	▪ Can understand talks on complex topics (e.g. environmental protection, public health) and distinguish main arguments from supporting evidence. ▪ Can understand academic lectures containing technical terms related to his/her own field and comprehend the main content. ▪ Can understand commentaries on current affairs, regardless of speech rate and accent; and evaluate main points. ▪ Can understand heated debates (e.g. court debates, public policy debates) regardless of speech rate; and evaluate the logic of the argumentation.
	Understanding oral interaction	▪ Can understand intentions and opinions of native English speakers who speak with a strong accent, when communicating with them on a wide range of topics. ▪ Can understand conversations using different English varieties (e.g. African American English, Indian English); and summarise main ideas. ▪ Can evaluate the rationality and logic of opinions of different sides when participating in impassioned academic discussions. ▪ Can understand group interviews regardless of speech rate and accent and identify the opinions of the interviewees.
CSE 7	Overall listening comprehension	▪ Can understand argumentation on abstract topics (e.g. politics, economy, history, culture); and evaluate the speakers' opinions and stance. ▪ Can follow interactions containing rhetorical devices (e.g. puns, metaphors) regardless of speech rate; and understand the speakers' implied meaning.
	Understanding oral description	▪ Can understand descriptions of natural landscapes (e.g. mountains, rivers) regardless of speech rate; and summarise their main features. ▪ Can understand detailed oral descriptions of buildings containing technical terms and obtain important information.
	Understanding oral narration	▪ Can understand news programs regardless of speech rate and comprehend the sociocultural connotations involved. ▪ Can understand live news reports on familiar topics regardless of background noise and summarise the main content. ▪ Can follow coverage on sports events and extract key information (e.g. athletes' performance, results).
	Understanding oral exposition	▪ Can understand complex introductory descriptions of operations, products, and services; and extract key information. ▪ Can understand advertisements containing slang or idioms, regardless of speech rate, and obtain specific information.
	Understanding oral instruction	▪ Can understand key points of multi-step technical instructions regardless of speech rate.
	Understanding oral argumentation	▪ Can understand speeches on abstract topics (e.g. politics, economy, culture) and evaluate speakers' opinions and stance. ▪ Can understand debates on public policies and social issues regardless of speech rate; and identify opinions and stance of both sides. ▪ Can understand academic conference presentations or debates in his/her own field and evaluate speakers' main points. ▪ Can understand academic discussions or talks in his/her own field and extract key concepts and main ideas.
	Understanding oral interaction	▪ Can obtain specific information from interactions with native English speakers regardless of speech rate, without asking for repetition or clarification. ▪ Can understand conversations containing puns and metaphors and infer speakers' implied meaning.
CSE 6	Overall listening comprehension	▪ Can understand highly-informative spoken discourse in his/her own field (e.g. lectures, presentations, discussions); summarise main ideas; and identify speakers' organisational patterns (e.g. overall framework, use of cohesive devices). ▪ Can understand common interactions in the workplace (e.g. business communications, job interviews) when produced at a normal speed; and identify the speakers' attitudes and intentions.
	Understanding oral description	▪ Can understand descriptions of certain places when delivered at a normal speed and grasp geographical features. ▪ Can understand subtle oral descriptions of emotional states of characters in stories and evaluate in relation to narrative development.

	Understanding oral narration	▪ Can understand radio programs delivered at a normal speed and identify speakers' opinions and stance. ▪ Can summarise the main content when watching TV programs delivered at a normal speed. ▪ Can understand news programs delivered at a normal speed and grasp main ideas. ▪ Can understand complex stories delivered at a normal speed and grasp the moral or philosophy.
	Understanding oral exposition	▪ Can understand open lectures in overseas universities in his/her own field and summarise the main content. ▪ Can understand coverage on the same event from different media when delivered at a normal speed and compare opinions from different sources. ▪ Can understand well-organised information about scenic spots when delivered with standard pronunciation; and obtain specific information.
	Understanding oral instruction	▪ Can understand broadcasts regardless of background noise in public places (e.g. stations, stadiums); and obtain specific information.
	Understanding oral argumentation	▪ Can understand highly-informative lectures or audio-taped/ video-taped talks and summarise key points and opinions. ▪ Can understand speeches on current affairs when delivered at a normal speed and summarise main points. ▪ Can follow conference presentations delivered with standard pronunciation at a normal speed and understand key points and specific details.
	Understanding oral interaction	▪ Can understand the cultural connotations of native English speakers' discourse when communicating with them on topics about daily life. ▪ Can understand conversations related to his/her own field and spoken with standard pronunciation; and distinguish primary from secondary information. ▪ Can understand conversations conducted with standard pronunciation and at a normal speed; and infer speakers' implied meaning. ▪ Can follow business negotiations articulated clearly and with standard pronunciation; and identify speakers' opinions and attitudes.
CSE 5	Overall listening comprehension	▪ Can understand spoken language on general topics when delivered at a normal speed; obtain main ideas and supporting details; identify logical relationships (e.g. causation, transition, progression); and understand the cultural connotations of expressions. ▪ Can understand radio, film, and TV programs on general topics and grasp main ideas.
	Understanding oral description	▪ Can obtain key information from descriptions of large-scale activities (e.g. festival celebrations, sports events), when delivered at a normal speed. ▪ Can understand native English speakers' oral descriptions of their social status quo and compare it with that of his/her own society.
	Understanding oral narration	▪ Can follow short news reports delivered at a normal speed and obtain factual information. ▪ Can understand humorous stories delivered at a normal speed and recognise the humour. ▪ Can understand complex novels delivered at a normal speed and identify personality traits of the main characters. ▪ Can understand familiar and popular English songs and grasp main ideas.
	Understanding oral exposition	▪ Can understand courses in his/her own field when delivered at a normal speed and note down key points. ▪ Can understand the main content of documentaries on familiar topics when delivered at a normal speed. ▪ Can understand introductory descriptions related to popular science when delivered at a normal speed; and grasp main ideas, provided topics are familiar.
	Understanding oral instruction	▪ Can understand multi-step instructions related to work and/or study. ▪ Can understand instructions for the use of everyday products when delivered with standard pronunciation at a normal speed.

	Understanding oral argumentation	▪ Can understand video programs or talks on general topics when delivered at a normal speed and obtain key points and details. ▪ Can understand speeches on social issues when delivered at a normal speed and distinguish opinions from facts. ▪ Can understand discussions on social issues when delivered at a normal speed and evaluate the logic of argumentation. ▪ Can understand academic talks delivered with standard pronunciation and identify speakers' organisational patterns (e.g. overall framework, use of cohesive devices). ▪ Can understand news commentaries on hot social issues and identify speakers' opinions and attitudes. ▪ Can understand television interviews on general topics when delivered at a normal speed and summarise key points and opinions.
	Understanding oral interaction	▪ Can understand conversations on hot social issues conducted at a normal speed and summarise main ideas. ▪ Can understand face-to-face interactions conducted at a normal speed and evaluate the appropriacy of speakers' expressions.
CSE 4	Overall listening comprehension	▪ Can understand spoken language delivered at a normal speed on general topics that are of personal interest (e.g. speeches, news reports, talks); distinguish primary from secondary information based on discourse features; and grasp the main idea. ▪ Can understand interactions on familiar topics and identify speakers' views and intentions.
	Understanding oral description	▪ Can follow simple oral descriptions of places of historical interest and obtain specific information (e.g. architectural style, geographical environment). ▪ Can understand descriptions of settings or people when delivered at a normal speed and obtain related information.
	Understanding oral narration	▪ Can understand and note down specific information when listening to radio programs or watching TV programs on general topics when delivered with standard pronunciation and at a normal speed. ▪ Can understand news reports delivered at a normal speed and identify causal relationships among events. ▪ Can follow the plot when listening to stories delivered at a normal speed. ▪ Can understand stories delivered at a normal speed and grasp the underlying message. ▪ Can understand anecdotes and travel logs delivered with standard pronunciation and at a normal speed and grasp main ideas. ▪ Can analyse relationships among characters and events when watching simple TV and film dramas.
	Understanding oral exposition	▪ Can follow information about a certain country or region when delivered at a normal speed; and obtain specific information (e.g. eating habits, customs, culture). ▪ Can understand weather reports delivered with standard pronunciation at a normal speed; and obtain specific information (e.g. regions, temperature, climatic features).
	Understanding oral instruction	▪ Can grasp main ideas of announcements and public notices when articulated clearly and delivered at a normal speed. ▪ Can understand general instructions (e.g. navigational directions, operation instructions) delivered at a normal speed.

	Understanding oral argumentation	▪ Can infer speakers' emotions and attitudes when listening to speeches on familiar topics or topics of personal interest. ▪ Can summarise main ideas with the help of pictures or videos, when listening to lectures on general topics delivered at a normal speed. ▪ Can understand debates on familiar topics articulated clearly and grasp the main arguments and supporting evidence on both sides. ▪ Can identify opinions of different sides and relationships among the opinions when listening to multi-party discussions on familiar topics. ▪ Can understand commentaries on familiar literary works or TV and film dramas; and infer speakers' emotions and attitudes.
	Understanding oral interaction	▪ Can understand conversations about daily life with standard pronunciation and at a normal speed; and identify the opinions and stance of different speakers. ▪ Can understand conversations conducted at a normal speed and evaluate the relevance between pieces of information and the topic. ▪ Can understand conversations conducted at a normal speed with subtleties in tone and infer speakers' implied meaning. ▪ Can understand interview questions delivered at a normal speed and identify interviewers' intentions.
CSE 3	Overall listening comprehension	▪ Can understand short speech (e.g. talks, discussions, announcements) delivered with standard pronunciation at a slow but natural speed; and obtain key information with the help of stress, intonation, background knowledge, and contextual information. ▪ Can identify themes and obtain main ideas when listening to radio or when watching film and TV programs on familiar topics and at a slow but natural speed.
	Understanding oral description	▪ Can follow simple oral descriptions of familiar countries or regions and obtain geographical location. ▪ Can follow simple descriptions of animals when delivered at a slow but natural speed and obtain information about physical features.
	Understanding oral narration	▪ Can understand simple stories or narratives delivered slowly but naturally and identify logical relationships among characters and events. ▪ Can follow accounts of personal experience when delivered with standard pronunciation and at a slow but natural speed; and obtain specific information such as time, place, and relationships among characters. ▪ Can follow radio programs on familiar topics when delivered slowly but naturally and identify the themes.
	Understanding oral exposition	▪ Can understand information about daily life (e.g. health and diet, safety knowledge); and grasp the main idea, provided speech is articulated clearly and delivered with standard pronunciation at a slow but natural speed. ▪ Can follow information about scenic spots in simple language when delivered at a slow but natural speed; and obtain specific information (e.g. historical, geographical). ▪ Can understand information about familiar products in simple language and identify key information, provided speech is articulated clearly and delivered at a slow but natural speed.
	Understanding oral instruction	▪ Can understand notices or multi-step instructions delivered at a slow but natural speed and grasp key points. ▪ Can understand broadcasts in public places (e.g. airports, stations) when delivered with standard pronunciation at a slow but natural speed; and obtain key information. ▪ Can follow explanations in simple language on the procedures for simple activities (e.g. handicrafts) when articulated clearly and delivered slowly.
	Understanding oral argumentation	▪ Can obtain key information from speeches or talks articulated clearly and delivered with standard pronunciation at a slow but natural speed. ▪ Can understand short argumentation on familiar topics that is delivered in simple language at a slow but natural speed; and grasp the main idea.
	Understanding oral interaction	▪ Can understand simple conversations in study and work and identify speakers' intentions. ▪ Can understand short conversations while shopping and obtain specific information (e.g. prices, sizes). ▪ Can follow formal conversations conducted at a slow but natural speed and identify topic progression and transition.

CSE 2	Overall listening comprehension	▪ Can understand speech (e.g. stories, talks, daily conversations) containing commonly used words delivered with standard pronunciation at a slow speed; and obtain information such as characters, time, places, and events.
	Understanding oral description	▪ Can follow simple oral descriptions of people, places, and common objects with the help of visuals and gestures; and obtain related information. ▪ Can follow simple oral descriptions of pictures and obtain specific information in the pictures (e.g. people, objects).
	Understanding oral narration	▪ Can understand narratives about daily life and personal information when delivered slowly and grasp main ideas. ▪ Can follow simple stories containing few low-frequency words when delivered at a slow speed; and obtain specific information (e.g. characters, places, events). ▪ Can understand simple stories containing few low-frequency words when delivered at a slow speed and infer causal relationships among events.
	Understanding oral exposition	▪ Can understand information about schedules in simple language articulated clearly and delivered slowly; and obtain specific information (e.g. time, places).
	Understanding oral instruction	▪ Can understand simple navigational directions articulated clearly and delivered slowly. ▪ Can understand simple instructions in handcraft making.
	Understanding oral argumentation	▪ Can understand short speeches or talks delivered with standard pronunciation at a slow speed and grasp main ideas.
	Understanding oral interaction	▪ Can follow short daily conversations and obtain specific information (e.g. places, events, relationships among people), provided speech is articulated clearly and delivered at a slow speed. ▪ Can follow short daily conversations and obtain numerical information (e.g. temperature, age, telephone numbers), provided speech is articulated clearly and delivered at a slow speed. ▪ Can follow telephone conversations about daily life, when conducted at a slow speed, and understand speakers' intentions.
CSE 1	Overall listening comprehension	▪ Can follow speech in simple language when articulated clearly and delivered slowly; identify words and phrases about oneself, his/her family, and familiar things; respond to simple instructions; and identify speakers' emotions and attitudes with the help of their stress, intonation, gestures, and facial expressions.
	Understanding oral description	▪ Can follow oral descriptions of common objects and identify specific information (e.g. colours, numbers). ▪ Can understand oral descriptions of a person that use simple words and identify who is being described.
	Understanding oral narration	▪ Can understand simple statements about common objects. ▪ Can understand main ideas of dialogues or monologues in cartoons that employ simple language. ▪ Can understand stories in simple language in picture books and match the words with pictures.
	Understanding oral exposition	▪ Can understand concrete information, such as information about family members, in simple language delivered slowly (e.g. "She is my mom"; "We are brothers").
	Understanding oral instruction	▪ Can understand simple instructions in daily life (e.g. cleaning rooms, opening or closing doors and/or windows). ▪ Can understand simple instructions in children's games.
	Understanding oral interaction	▪ Can understand simple greetings in daily life. ▪ Can follow simple conversations and identify words and phrases about oneself, his/her family, and school, provided speech is articulated clearly and delivered at a slow speed. ▪ Can follow short conversations conducted at a slow speed and identify numbers from 1 to 100 and common colours. ▪ Can understand simple questions about personal information (e.g. name, age, nationality).

A2 CSE Reading scales by levels

CSE level	Scales	Descriptor
CSE 9	Overall reading comprehension	▪ Can understand linguistically complex materials from a variety of fields, analysing them synthetically from multiple perspectives. ▪ Can synthetically appraise complex and abstruse specialised materials from relevant fields of study.
	Understanding written description	▪ Can appraise the value of linguistically complex descriptions drawn from a variety of fields and perspectives. ▪ Can appreciate the aesthetic qualities (e.g. those pertaining to language, ideas, or a realm) of lengthy prose essays written in complex and evocative language.
	Understanding written narration	▪ Can appraise the aesthetic value of poems which are written in an abstract manner and rich in imagery. ▪ Can understand conflicts and their significance in the script of a play with complex language and plot. ▪ Can make good sense of the value of a literary work with complex language and plot. ▪ Can appreciate the features and devices used in a literary work.
	Understanding written exposition	▪ Can appraise the applicability of linguistically complex expository writing in a highly specialised field. ▪ Can rapidly find relevant information in expository writing in a specialised field. ▪ Can summarise the main ideas of news stories about specialised fields. ▪ Can understand the main ideas set out in expository writing beyond his/her area of specialisation.
	Understanding written instruction	▪ Can appraise the use of language in instructions in specialised areas (e.g. instructions for software).
	Understanding written argumentation	▪ Can understand the conclusions of a research project outside his/her fields of study. ▪ Can make inductions from interdisciplinary academic monographs with profound content and abstract concepts. ▪ Can make a critical appraisal of the stance reflected in commentaries on social phenomena or topical issues.
	Understanding written interaction	▪ Can appraise the diction in diplomatic correspondence.
CSE 8	Overall reading comprehension	▪ Can discriminate and appreciate aesthetic language use and social significance of linguistically complex materials from a wide range of topics. ▪ Can appraise, by means of text analysis, the language and content of linguistically complex academic materials from familiar fields of study.
	Understanding written description	▪ Can recognise the author's intent in giving a particular description in a literary work. ▪ Can sum up the main claims of lengthy prose essays with philosophical content and complex language. ▪ Can appreciate the aesthetics of the language of linguistically complex articles on culture or art.

	Understanding written narration	▪ Can get the gist of a poem written in relatively complex language. ▪ Can summarise the main idea of a literary work with complex language and plot. ▪ Can evaluate the stylistic features of a linguistically complex novel.
	Understanding written exposition	▪ Can appraise the author's viewpoint in a linguistically complex article found in an English language newspaper or journal. ▪ Can summarise the main ideas in linguistically complex articles in political or economic newspapers and journals. ▪ Can analyse the relationship between different factors in charts in specialised fields. ▪ Can comprehend the implication of data in charts based on academic research.
	Understanding written instruction	▪ Can discern the implicit cultural differences in instructive texts used in different social contexts and cultural situations.
	Understanding written argumentation	▪ Can extract the core information from literature in relevant fields of study. ▪ Can make a critical analysis of the logic of arguments in linguistically complex argumentative writing. ▪ Can understand the aims, methods, and conclusions of an academic paper in relevant fields of study. ▪ Can obtain information needed for research from academic monographs in relevant fields of study. ▪ Can make a comparative study of the research methods used in the literature in relevant fields of study. ▪ Can differentiate opinions from facts in commentaries written in complex language and aimed to answer incisive questions. ▪ Can make judgements on the contributions and limitations of academic papers in relevant fields of study. ▪ Can evaluate the rationality of viewpoints in book reviews in relevant fields of study.
	Understanding written interaction	▪ Can discern the social-historical traits in substantial correspondence between historical figures.
CSE 7	Overall reading comprehension	▪ Can synthesise the content of specialised linguistically complex materials (e.g. original literary works, science and technology literature, social commentaries), and analyse the author's viewpoint and stance. ▪ Can make critical comments on a variety of cultural phenomena from different cultures, as presented in linguistically complex works. ▪ Can comprehend the implicit meaning of specialised linguistically complex materials by relating the materials to similar topics.
	Understanding written description	▪ Can understand the cultural connotations of linguistically complex articles on social culture. ▪ Can appreciate the linguistic features of prose essays written in relatively complex language. ▪ Can analyse, from specific perspectives, the descriptive methods in different articles concerning the same social phenomenon. ▪ Can analyse the language of linguistically complex lyrical prose essays. ▪ Can summarise information in connection with people's feelings to explore the inner world of characters presented in lengthy, linguistically complex prose essays. ▪ Can analyse linguistically complex descriptive articles in order to infer the author's stance on significant topics (e.g. national spirit, social ideals). ▪ Can appreciate rhetorical devices in lengthy, linguistically complex lyrical prose essays.

	Understanding written narration	▪ Can infer the intended meaning of a short story with relatively complex language and plot. ▪ Can appraise the style of a play script with complex language and plot. ▪ Can appraise the narrative methods used in biographical writing with relatively complex language. ▪ Can sum up the typical cultural characteristics in a work about culture and written in relatively complex language. ▪ Can understand the gist of excerpts from literary classics, including fiction and drama, written in complex language.
	Understanding written exposition	▪ Can summarise the main ideas of specialised reports containing numerous pictures and graphs. ▪ Can discern the key information in linguistically complex articles on science and technology of long or average length. ▪ Can summarise the key information in linguistically complex expository writing containing jargon. ▪ Can summarise the main features of descriptions in linguistically complex expository writing on social sciences. ▪ Can make a comparative study of the traits of different cultures around the world as they are described in linguistically complex expository writing. ▪ Can understand research methods as described in research reports written in relatively complex language in specialised fields. ▪ Can understand the meaning of the data in charts in specialised fields.
	Understanding written instruction	▪ Can understand the requirements of instructive texts such as professional manuals. ▪ Can appraise the diction in explanatory writing on laws and regulations.
	Understanding written argumentation	▪ Can appreciate the features of the language used in linguistically complex argumentative writing. ▪ Can extract the main information from literature pertinent to relevant fields of study. ▪ Can infer the author's feelings and attitudes in argumentative texts on profound topics. ▪ Can get the main idea of book reviews in relevant fields of study. ▪ Can understand the logic of the author's thought in a linguistically complex argumentative text. ▪ Can appraise the effectiveness of arguments in linguistically complex argumentative texts. ▪ Can evaluate the practical significance of commentaries on social issues.
	Understanding written interaction	▪ Can discover evidence of the author's viewpoints and feelings in linguistically complex and long letters. ▪ Can appraise the diction in government documents written in relatively complex language. ▪ Can appraise the appropriacy of diction in business correspondence written in relatively complex language. ▪ Can make a comparative analysis of the diction and style used in a variety of correspondence.
CSE 6	Overall reading comprehension	▪ Can grasp significant relevant information and briefly comment on the language and content of subject-related materials of medium linguistic difficulty (e.g. literary works, news reports, business documents). ▪ Can infer the writer's mood and attitude while reading materials of medium linguistic difficulty (e.g. literary works, news reports). ▪ Can locate target information by scanning the indices of academic literature.
	Understanding written description	▪ Can infer the author's attitude in a medium-length descriptive article written in relatively complex language. ▪ Can summarise the major features of a scene described in an essay written in relatively complex language. ▪ Can evaluate the descriptive methods used in descriptions of people, events, or objects written in relatively complex language.

	Understanding written narration	▪ Can analyse the narrative methods of stories which are relatively complex in terms of language and plot. ▪ Can locate the significant details of literary works such as novels which are written in relatively complex language. ▪ Can comprehend the implicit meaning in popular editions of works of philosophy. ▪ Can comprehensively understand characters or events pertaining to social life in stories written in relatively complex language. ▪ Can understand the logical progression in excerpts from novels written in relatively complex language. Can understand the implicit meanings and allusions in fairy tales.
	Understanding written exposition	▪ Can summarise the main claims made in popular science articles written in relatively complex language. ▪ Can identify the key points of information supplied by common news stories. ▪ Can summarise the main ideas of expository texts about the development of science and technology written in relatively complex language. ▪ Can summarise the major traits of folk customs as described in expository texts written in relatively complex language. ▪ Can understand the major findings of surveys on social life written in complex language.
	Understanding written instruction	▪ Can understand the terminology of operational texts in related professional areas. ▪ Can extract detailed information from instructive writing. ▪ Can understand the instructions in instructive writing.
	Understanding written argumentation	▪ Can comprehend core information from speeches or reports written in relatively complex language. ▪ Can understand the modes of argumentation employed in argumentative texts on social phenomena expressed in relatively complex language. ▪ Can understand and summarise the author's viewpoints and stances in commentaries written in relatively complex language. ▪ Can conduct a critical analysis of the persuasiveness of argumentative texts written in relatively complex language. ▪ Can generalise the main ideas of literature reviews in relevant disciplines. ▪ Can appraise the effectiveness of language used in argumentative texts written in relatively complex language. ▪ Can evaluate the logic of speeches written in relatively complex language.
	Understanding written interaction	▪ Can appraise the diction in commercial correspondence. ▪ Can summarise the main ideas of relatively long situational dialogues about social life.
CSE 5	Overall reading comprehension	▪ Can grasp essential meaning, analyse linguistic features, and understand cultural implications whilst reading materials of medium linguistic difficulty on a variety of topics likely to be encountered in the domains of education, technology, and culture. ▪ Can distinguish different positions in materials of medium linguistic difficulty containing opposing argumentation (e.g. editorials, book reviews).
	Understanding written description	▪ Can make a comparative analysis of descriptive methods used in essays on the same topic. ▪ Can understand the author's feelings in short prose essays written in relatively complex language. ▪ Can comprehend the rhetorical devices found in linguistically simple prose essays on social life. ▪ Can understand linguistically simple essays about a character's mentality to comprehend the character's personality or change in emotion.
	Understanding written narration	▪ Can understand figures of speech (e.g. tropes and personifications) in stories written in relatively complex language. ▪ Can understand the emotion contained in a poem written in simple language. ▪ Can extract relevant details from social life stories which are relatively complex in language and plot. ▪ Can summarise the major plot details of a novel written in simple language. ▪ Can understand the relationship between elements such as characters and events in narrative prose essays written in relatively complex language.

	Understanding written exposition	▪ Can analyse the linguistic features of popular science articles written in relatively complex language. ▪ Can extract the key information in practical forms of writing (e.g. memos or notes) written in relatively complex language. ▪ Can understand an expository text's description of a product and its use in everyday life. ▪ Can analyse data trends presented in common charts. ▪ Can comprehend the key points made in news with captions explaining social phenomena or topical issues written in relatively complex language.
	Understanding written instruction	▪ Can understand the specific technical requirements of writing in related professional areas.
	Understanding written argumentation	▪ Can analyse argumentative texts on common topics to infer the authors' implicit viewpoints and attitudes. ▪ Can conduct a comparative or contrastive analysis of different authors' viewpoints on the same topic in argumentative texts. ▪ Can summarise the viewpoints and arguments in commentaries on familiar topics. ▪ Can understand the connection between the viewpoints and illustrations in argumentative texts on topical social issues. ▪ Can identify the main ideas of an author's review of a non-academic book. ▪ Can understand the gist of a commentary on social phenomena or topical issues. ▪ Can distinguish between opinions and facts when reading speeches on familiar topics.
	Understanding written interaction	▪ Can extract the core information from practical forms of writing (e.g. letters of application and recommendation). ▪ Can understand the main points in simple business dialogues. ▪ Can comprehend the author's stance in material about social phenomena (e.g. letters and blogs).
CSE 4	Overall reading comprehension	▪ Can locate detailed information and summarise the main idea whilst reading different kinds of linguistically simple materials (e.g. simple short stories, essays, letters). ▪ Can differentiate facts and opinions and make simple inferences in linguistically simple narratives and argumentative texts on a variety of topics. ▪ Can understand the relationship between ideas by analysing the structures of sentences and discourse whilst reading materials of medium linguistic difficulty.
	Understanding written description	▪ Can specify the major features of a scene described in linguistically simple travel writing. ▪ Can identify the way an author expresses himself/herself in linguistically simple lyrical prose.
	Understanding written narration	▪ Can recognise details (e.g. time, character, and place) in articles on social life, such as travel notes, written in relatively complex language. ▪ Can extract details that communicate a character's feelings and attitudes from stories written in relatively complex language. ▪ Can work out a character's personality from anecdotes written in relatively complex language. ▪ Can work out the progression of historical stories which are relatively complex in terms of language and plot. ▪ Can understand the author's stance or viewpoint in short articles about daily life, written in relatively complex language. ▪ Can understand the main idea of a biography with simple language and plot. ▪ Can infer the author's intent from a narrative of social life written in simple language. ▪ Can distinguish between primary and secondary plots in excerpts from novels written in simple language.

	Understanding written exposition	▪ Can understand the main points made in short popular science articles. ▪ Can summarise the main points made in short expository essays on Chinese and foreign cultures. ▪ Can understand the meaning of data in simple charts. ▪ Can understand the main points made in notices, posters, and advertisements in everyday life. ▪ Can extract the key information from news stories on topical issues written in relatively complex language.
	Understanding written instruction	▪ Can understand the connections between steps in procedures in simple flow charts.
	Understanding written argumentation	▪ Can find the key words embodying the authors' viewpoints on topical issues in short, simple argumentative texts. ▪ Can analyse the authors' viewpoints on familiar social phenomena in short, simple argumentative texts. ▪ Can make judgements about the consistency between viewpoints and arguments in linguistically simple argumentative texts on topical issues. ▪ Can differentiate primary viewpoints from secondary ones in short argumentative texts on current affairs. ▪ Can distinguish between opinions and facts when reading linguistically simple speeches. ▪ Can understand the core ideas in linguistically simple philosophical argumentative texts.
	Understanding written interaction	▪ Can extract the key information from letters about everyday life. ▪ Can make a judgement about the appropriacy of the language used in linguistically simple letters. ▪ Can extract the core information from commercial correspondence.
CSE 3	Overall reading comprehension	▪ Can locate key information in linguistically simple practical forms of writing (e.g. letters, notices, signs). ▪ Can understand the implicit meaning and summarise the main points of short, linguistically simple materials on familiar topics. ▪ Can understand the relationship between points of information with the help of connectors in linguistically simple argumentative texts on familiar topics.
	Understanding written description	▪ Can understand the main features of customs or cultures as described in a short, linguistically simple essay. ▪ Can extract the main information about a scenic spot from a description written in simple language. ▪ Can extract key details about a scenic spot as described in a short, linguistically simple essay.
	Understanding written narration	▪ Can infer the implicit message of an anecdote written in simple language. ▪ Can extract the main outline of a historical story written in simple language. ▪ Can analyse the personalities of heroes and antiheroes in a story written in simple language. ▪ Can pick out the significant events in an abridged version of a biography. ▪ Can understand the implicit meaning of a fable in simple language. ▪ Can extract the essential meaning of paragraphs from a story about social life. ▪ Can understand the intent of the speakers when reading dialogues about daily life.
	Understanding written exposition	▪ Can spot specific information such as details pertaining to times and places in common practical forms of writing (e.g. notices or bulletins). ▪ Can recognise the core information in short, linguistically simple news stories on topical issues. ▪ Can recognise the key information in short, linguistically simple popular science articles. ▪ Can summarise the main claims in short, linguistically simple expository essays on culture.
	Understanding written instruction	▪ Can understand the task for each procedure in linguistically simple manuals. ▪ Can summarise the key content of linguistically simple instructive writing on technical requirements.

	Understanding written argumentation	▪ Can understand the gist of short, simple arguments about social life.
	Understanding written interaction	▪ Can extract the core information from official invitations. ▪ Can infer the intent of the speakers in brief and simple dialogues about everyday life. ▪ Can understand the authors' viewpoints in short, simple letters about social issues. ▪ Can distinguish primary from secondary information in short, simple notices or posters.
CSE 2	Overall reading comprehension	▪ Can acquire specific information and understand the main idea of short, linguistically simple essays on familiar topics. ▪ Can understand short, simple texts containing new words with the help of pictures or other methods.
	Understanding written description	▪ Can recognise the major conditions as described in a short, linguistically simple essay about an event. ▪ Can understand the main features of a person in a short, linguistically simple essay about a person.
	Understanding written narration	▪ Can list key elements, such as people, places, and events, in simple stories. ▪ Can categorise words pertaining to weather, colour, or animals in linguistically simple materials (e.g. children's songs and nursery rhymes). ▪ Can analyse the relationship between characters in short, simple everyday dialogues. ▪ Can recognise key words communicating the author's feelings in diary entries written in simple language. ▪ Can extract salient information pertaining to the main characters from anecdotes written in simple language. ▪ Can briefly summarise a character's personality as expressed in anecdotes written in simple language. ▪ Can grasp the sequence of events in simple stories.
	Understanding written exposition	▪ Can spot the key information in short, linguistically simple expository essays and deictic expressions about everyday life. ▪ Can understand the main claims in short, linguistically simple expository texts with plenty of illustrations.
	Understanding written instruction	▪ Can understand the major task for each procedure in linguistically simple instructions. ▪ Can understand public signs (e.g. traffic signs).
	Understanding written interaction	▪ Can pick out the key information in notes or notices in everyday life.
CSE 1	Overall reading comprehension	▪ Can understand very short, simple texts and locate basic information (e.g. characters, time, place). ▪ Can understand simple materials (e.g. children's songs and nursery rhymes) and identify common words.
	Understanding written narration	▪ Can feel the rhyme in children's songs or nursery rhymes. ▪ Can understand the main ideas in a simple picture book. ▪ Can distinguish characters from each other in a simple story. ▪ Can pick out and understand the common words in children's songs or nursery rhymes.

A3 CSE Speaking scales by levels

CSE level	Scales	Descriptor
CSE 9	Overall oral expression	▪ Can communicate extensively over a wide range of social and cultural topics and can adjust the content and manner freely and effectively. ▪ Can effectively communicate and negotiate complicated and controversial professional topics on formal occasions.
	Oral description	▪ Can accurately and smoothly describe the details of social problems in order to solve them (e.g. customer complaints and accident disputes).
	Oral narration	▪ Can tell stories from history, tales from literature, or anecdotes, with lively and vivid language and rich details, skilfully involving listeners.
	Oral exposition	▪ Can present his/her own research in detail and respond to questions coherently and logically at international conferences in his/her field. ▪ Can elaborate on abstract and complex issues, such as national policy, principles, and systems.
	Oral instruction	▪ Can give clear and well-constructed oral instructions in a style appropriate to the context and in a tone that would help the recipient to notice significant points.
	Oral argumentation	▪ Can give inspiring impromptu speeches on professional topics. ▪ Can articulate his/her views on foreign affairs and can effectively defend his/her position.
	Oral interaction	▪ Can smoothly and comfortably discuss abstract and complex social topics without any language barrier. ▪ Can effectively communicate and consult on complex and controversial issues in his/her field of expertise. ▪ Can use language flexibly and fluently when hosting events or being interviewed.
CSE 8	Overall oral expression	▪ Can effectively discuss a wide range of topics in formal and informal settings, using appropriate rhetorical devices to enhance the effect utterances. ▪ Can express his/her viewpoints accurately and fluently on professional topics at academic seminars. ▪ Can thoroughly and effectively communicate or consult on complex and controversial issues encountered at work.
	Oral description	▪ Can describe subtle differences between objects or emotions of a similar nature using appropriate vocabulary. ▪ Can vividly describe real or historical persons, locations, or events.
	Oral narration	▪ Can briefly and concisely re-tell an event reported by the media and elaborate on the details, if necessary.
	Oral exposition	▪ Can give a detailed explanation and interpretation of articles or speeches with relatively abstract content. ▪ Can explain and analyse the lines of argument in public speeches in a comprehensive and logical manner. ▪ Can make a complete and logical summary of the information in a video or audio recording. ▪ Can interpret or explain complex issues using logical analysis, such as identifying priorities and highlighting essential points. ▪ Can offer a lucid explanation of abstract theories. ▪ Can accurately convey the main ideas and supporting details in an academic lecture.
	Oral instruction	▪ Can give coherent and clear oral instructions on on-going experiments or research.
	Oral argumentation	▪ Can give speeches on professional topics and use appropriate evidence to support his/her arguments. ▪ Can give well-structured and logical argumentative speeches on familiar topics. ▪ Can discuss complicated and abstract topics and can express clear viewpoints and profound content.

	Oral interaction	▪ Can communicate smoothly with others in professional seminars. ▪ Can effortlessly participate in group discussions and debates on abstract and complex topics. ▪ Can participate fully and effectively in discussions on a wide range of topics in formal and informal settings. ▪ Can appropriately express personal opinions on challenging and sensitive topics. ▪ Can effortlessly and easily engage in debate about topics related to his/her professional background. ▪ Can communicate and negotiate effectively in business communication on topics such as business arrangements, product prices, and related matters. ▪ Can effectively negotiate compensation, liability, and other matters when dealing with disputes.
CSE 7	Overall oral expression	▪ Can discuss a variety of familiar topics, appropriately express the desire to speak, and hold the floor. ▪ Can express personal opinions about abstract topics and adjust the content and styles of expression. ▪ Can make formal academic presentations and provide further explanation based on questions, using accurate, clear, and coherent language.
	Oral description	▪ Can clearly and accurately describe his/her common symptoms when seeing a doctor.
	Oral narration	▪ Can, with preparation, narrate the allusions and legends about places of interest in detail.
	Oral exposition	▪ Can give a detailed and coherent report of the research that he/she is undertaking, such as reporting a project's progress or current priorities. ▪ Can make impromptu speeches in a coherent and logical manner on topics related to school, work, and the community. ▪ Can elaborate on his/her plan or action in a persuasive manner. ▪ Can give a comprehensive and accurate summary of the materials that he/she has read. ▪ Can briefly analyse literary reviews, including the choice of diction, delivery, and effect. ▪ Can make relatively in-depth oral comments on literary works, movies, TV programs, or artistic works. ▪ Can briefly explain the artistic effects of literary works, movies, and TV programs and explain the techniques used in their creation. ▪ Can give a detailed explanation of topics in his/her own field in a logical and comprehensible manner. ▪ Can give a coherent oral report of the procedure and result of an experiment or investigation.
	Oral instruction	▪ Can give clear, explicit, and detailed oral instructions on how to handle complex affairs (e.g. business negotiations or symposium programs).
	Oral argumentation	▪ Can extensively and coherently elaborate on his/her views on academic or professional topics. ▪ Can express his/her own viewpoints on social or cultural topics and provide extensive supporting arguments. ▪ Can give persuasive speeches in speech contests. ▪ Can organise ideas logically on formal occasions, such as seminars in his/her field, and can select appropriate evidence to support them. ▪ Can provide convincing arguments in speeches based on specific topics by appropriately using statistics, presenting evidence, citing examples, and incorporating other means. ▪ Can make appropriate comments on others' views, inferences, and argumentation in discussion. ▪ Can synthesise and refine the main points in professional discussions to demonstrate comprehensive understanding of the problems under discussion.
	Oral interaction	▪ Can converse smoothly and thoroughly with others on familiar topics. ▪ Can use persuasive language to lodge compensation claims and indicate the bottom line in dealing with disputes. ▪ Can discuss medical procedures or speak with medical staff during medical treatments. ▪ Can communicate spontaneously with others on popular social issues and express views clearly and logically. ▪ Can confidently respond to questions during an interview.

CSE 6	Overall oral expression	▪ Can discuss hot social issues or familiar topics in his/her field and respond appropriately to remarks, interruptions, etc. ▪ Can make some insightful remarks on given topics related to social culture and learning; has a broad repertoire of oral expressions and can speak in a coherent and organised manner. ▪ Can communicate or negotiate effectively when dealing with daily disputes or unexpected situations.
	Oral description	▪ Can give a detailed and accurate description of scenery or places. ▪ Can describe in detail the main characters and scenes of a story or a movie. ▪ Can give a detailed and orderly description of the progression of hot social issues. ▪ Can describe his/her plans or experiences using a rich and accurate vocabulary.
	Oral narration	▪ Can paraphrase the content of an article in a complete and detailed manner, if the topic is familiar. ▪ Can vividly describe the events he/she has experienced in detail. ▪ Can describe an emergency incident in detail, such as reporting the entire course of an event to the police. ▪ Can adapt or continue a story using his/her own words.
	Oral exposition	▪ Can present a detailed analysis with comments of an article or an interview in a coherent and logical manner. ▪ Can clearly interpret and analyse the current state of, reasons for, and solutions to a hot social issue. ▪ Can give an explanation of the causes of a problem and propose a solution during a discussion. ▪ Can briefly present and explain viewpoints in formal meetings or seminars. ▪ Can elaborate on relevant details based on communicative needs.
	Oral instruction	▪ Can give clear and explicit oral instructions on the methods, steps, or procedures for implementing certain plans (e.g. a spring outing or making a donation). ▪ Can deliver detailed instructions or make requests to classmates/colleagues at school/work.
	Oral argumentation	▪ Can adequately express opinions and articulate his/her stance with sufficient evidence on hot social issues, such as environment protection. ▪ Can effectively persuade others to adopt her/his views on education, careers, and other life choices, by analysing the situation from multiple perspectives and weighing the advantages and disadvantages. ▪ Can express personal opinions with reasonable insight on given topics. ▪ Can give in-depth comments on the information he/she has heard, seen, or read. ▪ Can compare reports on an event that has appeared in news media such as TV and the internet and express opinions with reasonable insight.
	Oral interaction	▪ Can communicate smoothly in daily business settings. ▪ Can communicate politely and appropriately with others during disputes that arise in everyday life, study, or work settings. ▪ Can effectively debate popular social issues. ▪ Can respond appropriately to unexpected comments, such as criticisms and queries in everyday life.

CSE 5	Overall oral expression	▪ Can comment or communicate on everyday topics, as well as familiar topics and popular social issues, using clear, organised, and logical language. ▪ Can, after preparation, briefly comment on topics in his/her field. ▪ Can communicate or consult effectively on matters of daily life, such as business, travel, and shopping.
	Oral description	▪ Can briefly compare different cultural traditions or customs (e.g. Chinese and Western festivals). ▪ Can accurately describe his/her feelings (e.g. joy, sorrow, or fear). ▪ Can describe in detail common events/objects, character traits, and everyday scenes. ▪ Can describe, in an orderly way, arrangements for activities (e.g. class meetings or family gatherings).
	Oral narration	▪ Can describe his/her personal experiences at different stages of life in detail, such as education or life experience. ▪ Can re-tell the main idea of a story he/she has read, using the words or sentences of the original text. ▪ Can describe the main plot of a novel or film.
	Oral exposition	▪ Can give a coherent and detailed explanation of his/her choices or personal preferences and give justifications. ▪ Can give a clear and coherent explanation of a plan, such as what to do and how to do it. ▪ Can effectively communicate with others on his/her circumstances and needs in the case of an emergency, such as reporting an emergency or making an emergency call. ▪ Can, after consulting operation manuals or guidelines, give a brief explanation of procedures or policies such as operating a computer or applying for a driver's license. ▪ Can, after preparation, present views coherently on hot social issues. ▪ Can briefly explain the basic rules of some popular sports or games.
	Oral instruction	▪ Can present detailed travel plans to others with the help of a map or a guidebook. ▪ Can explicitly explain the operational system of an instrument or device by referring to a manual.
	Oral argumentation	▪ Can compare the advantages and disadvantages of different choices and make decisions accordingly, such as where to travel or which products to buy. ▪ Can, after preparation, express brief opinions on topics relevant to his/her background. ▪ Can give brief comments on literary works, such as movies, calligraphy, paintings, and novels. ▪ Can present ideas logically and effectively emphasise main points in a speech. ▪ Can give pertinent comments on other people's speeches. ▪ Can clearly express his/her views on and attitudes towards hot topics under discussion. ▪ Can elaborate on his/her views with concrete examples in formal speeches.
	Oral interaction	▪ Can effectively consult with relevant staff on study and work issues. ▪ Can engage in discussions with others on popular social issues. ▪ Can answer unprepared questions in interviews or after a presentation. ▪ Can carry out a prepared interview, checking and confirming respondents' information. ▪ Can conduct simple negotiations when problems arise on a trip. ▪ Can communicate effectively with others on familiar topics during audio or video conferences. ▪ Can briefly discuss with peers an article that he/she has read. ▪ Can participate in academic discussions and communications using simple language. ▪ Can engage in simple conversations about routine maintenance services (e.g. repair and maintenance of equipment or replacement of facilities). ▪ Can communicate effectively with salespeople when shopping (e.g. asking for information about items and bargaining).

CSE 4	Overall oral expression	▪ Can convey personal needs and wishes using appropriate expressions, such as different degrees of politeness. ▪ Can communicate on topics of interest and engage in smooth communication with timely responses. ▪ Can relate personal experiences or short stories in a coherent manner. ▪ Can briefly describe or explain common activities or scenes in daily life and work, such as sports, recreational activities, and scenic spots.
	Oral description	▪ Can briefly describe audio or video recordings on familiar themes. ▪ Can briefly describe his/her school or workplace (e.g. its location or occupants). ▪ Can describe in detail his/her wishes or aspirations (e.g. expected trips or ideal jobs). ▪ Can introduce him/herself in detail (e.g. study, work, hobbies). ▪ Can briefly describe past or upcoming events in chronological order. ▪ Can briefly describe cultural traditions or customs. ▪ Can briefly describe his/her symptoms when seeing a doctor.
	Oral narration	▪ Can tell a short story in a relatively complete and coherent manner. ▪ Can retell the main plot of a short story in a relatively complete manner. ▪ Can briefly report what happened during an event. ▪ Can coherently describe a personal experience, such as a trip.
	Oral exposition	▪ Can give short introductory descriptions for courses available at a school after consulting course descriptions or other relevant materials. ▪ Can give a detailed description of his/her life plans and explain the reasons. ▪ Can give a detailed explanation of everyday occurrences, such as running late or being absent. ▪ Can briefly describe why he/she is qualified for a job or position. ▪ Can briefly describe the steps of common activities in work or study contexts. ▪ Can give a brief description of, or commentary on, a famous person, a place of historical interest, or a cultural tradition. ▪ Can, after preparation, give a short presentation on topics related to school, society, or work. ▪ Can, after preparation, elaborate on the reasons he/she likes a particular movie or a piece of music. ▪ Can, after preparation, briefly present views on social problems, such as pocket money, the generation gap, and being rebellious.
	Oral instruction	▪ Can briefly respond to others' enquiries in daily communication (e.g. asking for directions). ▪ Can give simple oral instructions during familiar activities, such as sports or games. ▪ Can outline the operational procedures of common electrical appliances (e.g. computers and smart phones).
	Oral argumentation	▪ Can give brief oral comments on a written text. ▪ Can comment on a presentation given by others. ▪ Can explain the main points in his/her speech and support them with appropriate evidence.
	Oral interaction	▪ Can engage in simple conversations about goods information (e.g. colour, size, style, price of goods) when shopping. ▪ Can briefly discuss topics such as family or school. ▪ Can discuss his/her dreams or plans with others, such as plans for studying abroad or careers he/she would like to pursue. ▪ Can conduct brief conversations about exchanging goods, refunds, or other business matters. ▪ Can make requests and carry on simple negotiations about services, bills, and other transactions. ▪ Can book everyday services over the telephone (e.g. medical appointments, tickets, meals). ▪ Can ask about accommodation and travel arrangements when receiving foreign guests. ▪ Can briefly answer customers' inquiries at work. ▪ Can talk to staff at a rental company about hiring vehicles. ▪ Can speak with clerks at the bank about day-to-day tasks (e.g. opening or cancelling accounts, making deposits or withdrawals). ▪ Can conduct day-to-day communications about everyday learning, such as arranging appointments or asking for course/exam information.

CSE 3	Overall oral expression	▪ Can communicate briefly on familiar topics with fair coherence and reasonable accuracy in pronunciation, intonation, and tense. ▪ Can, with the help of others, participate in discussions in professional or academic settings. ▪ Can communicate with simple expressions, using strategies such as paraphrasing and word coinage for effective communication when necessary.
	Oral description	▪ Can, with the help of others, describe his/her or other people's experiences in simple terms prompted by a picture or text. ▪ Can express personal needs, wishes, and feelings in simple terms. ▪ Can describe common Chinese and international festivals in simple terms.
	Oral narration	▪ Can briefly describe the elements of an event, such as time, place, and character. ▪ Can briefly re-tell or paraphrase what others have said. ▪ Can briefly re-tell the main idea of a short passage.
	Oral exposition	▪ Can, after preparation, use simple words to express views or ideas on topics that are pertinent to his/her everyday life or study. ▪ Can accurately convey information given by teachers, such as homework, lesson planning, and class schedules.
	Oral instruction	▪ Can give brief, short, and common instructions or commands in oral interactions. ▪ Can give short oral instructions to assign roles and tasks during tasks involving group cooperation.
	Oral argumentation	▪ Can use simple language to express his/her views on everyday topics. ▪ Can give short responses to inquiries from others. ▪ Can cite examples from everyday occurrences as evidence to support his/her viewpoint.
	Oral interaction	▪ Can participate in group discussions about related content using simple language. ▪ Can exchange information with others on familiar topics. ▪ Can participate in group discussions about everyday topics with the help of others. ▪ Can communicate with others, such as hotel front desk staff, for general needs and inquiries.
CSE 2	Overall oral expression	▪ Can handle short everyday exchanges with simple terms and clear pronunciation, using intonation appropriate and natural for the situation. ▪ Can give a short, simple presentation or narration after preparation, resorting to strategies such as lexical substitution to compensate for linguistic deficiency. ▪ Can express his/her views with the help of pictures or other people, such as familiar people, objects, and places.
	Oral description	▪ Can describe, after preparation, his/her dreams or aspirations in simple terms. ▪ Can describe his/her feelings in simple terms (e.g. happiness, anger, excitement, boredom). ▪ Can describe the features of familiar objects in simple terms (e.g. material, weight, shape). ▪ Can describe the characteristics of familiar persons in simple terms (e.g. appearance, clothes, occupation). ▪ Can describe familiar places in simple terms (e.g. hometown and cities visited). ▪ Can describe his/her recent work or studies in simple terms. ▪ Can describe the role he/she assumes or plays in simple terms in either real-life or simulated situations.
	Oral narration	▪ Can, with preparation, tell simple and familiar short stories. ▪ Can create and role-play short dialogues; or tell short stories and narrate personal experiences with the support of others or pictures.
	Oral exposition	▪ Can, with help from others, use simple words to describe functions of everyday items. ▪ Can use simple words to explain common rules, regulations, or instructions, such as basic traffic rules or indication signs.

	Oral instruction	▪ Can provide simple information about routes, directions, and other topics in familiar communication situations. ▪ Can provide simple oral instructions, such as “Open the door” and “Stand up”.
	Oral argumentation	▪ Can use simple language to express his/her own opinions, based on the provided verbal cues or with the help of others.
	Oral interaction	▪ Can ask about others’ feelings in everyday conversation. ▪ Can talk with others and decide on the time and place for get-togethers or meetings. ▪ Can negotiate time arrangements with others using simple words and phrases. ▪ Can describe dietary requirements using simple terms and ask about prices when ordering a meal.
CSE 1	Overall oral expression	▪ Can name common objects. ▪ Can express personal preferences and introduce him/herself or familiar people in simple terms, using body language or demonstrative pronouns when necessary. ▪ Can participate in routine communicative activities when support is provided and use simple terms to ask for repetition in case of incomprehension.
	Oral description	▪ Can describe his/her habits and routines in simple terms. ▪ Can give basic information about him/herself (e.g. name, age, birthplace). ▪ Can describe the length, size, and colour of everyday objects in simple terms.
	Oral narration	▪ Can name objects in pictures. ▪ Can read short dialogues.
	Oral exposition	▪ Can use simple words to express his/her likes and dislikes.
	Oral instruction	▪ Can ask and answer simple questions about the position of certain objects.
	Oral argumentation	▪ Can express his/her attitudes to familiar events or actions, such as agreement or disagreement and approval or disapproval.
	Oral interaction	▪ Can exchange greetings or holiday wishes with others.

A4 CSE Writing scales by levels

CSE level	Scales	Descriptor
CSE 9	Overall written expression	▪ Can write critical articles on, or critiques pertaining to, a range of topics with a comprehensive consideration of writing purpose, readership, and communicative settings. ▪ Can produce literary creations based on social phenomena, using expressive language full of artistic appeal and with distinctive features of style and register.
	Written description	▪ Can describe natural scenery from different angles with appropriate, vivid, and visual language. ▪ Can vividly and meticulously describe the psychological features that reveal the personality of a character in creative writing.
	Written narration	▪ Can produce creative work based on social phenomena, using authentic, natural, and elegant linguistic expression.
	Written exposition	▪ Can write detailed product specifications in language conforming to industry norms.
	Written argumentation	▪ Can write responses to complex viewpoints in academic journal articles, marshalling sufficient evidence and giving convincing arguments. ▪ Can write editorials for newspapers and/or magazines, articulating clear viewpoints and convincing arguments on social topical issues.
	Written interaction	▪ Can compose formal and standard contracts based on given content. ▪ Can compose international co-operation agreements or treaties for departments of foreign, commercial, or other affairs.
CSE 8	Overall written expression	▪ Can discuss complex social problems with clarity, organisation, and logic. ▪ Can appropriately summarise and evaluate relevant literature and write academic articles with sufficient evidence, in-depth discussion and reliable conclusion(s). ▪ Can produce some creative writing with fluent language, relatively good structure, and depth.
	Written description	▪ Can objectively describe and analyse the salient features of a series of events. ▪ Can concisely describe features of Chinese culture to foreigners in an appropriate and coherent manner.
	Written narration	▪ Can produce creative work in certain genres (e.g. skilfully-designed humorous short stories).
	Written exposition	▪ Can write substantive reports on topics related to history, society, or culture. ▪ Can use easy-to-understand language to explain professional knowledge unfamiliar to readers. ▪ Can explain certain abstract concepts in a clear and logical way.
	Written instruction	▪ Can write clear and well-organised research procedures and methods for projects in his/her discipline.
	Written argumentation	▪ Can provide comprehensive summaries of and commentaries on viewpoints presented in broadcasted interviews or debates. ▪ Can analyse the causes of complex social problems and develop a sound argument. ▪ Can write literature reviews based on critical reading, with a clear structure and legitimate analysis.
	Written interaction	▪ Can compose formal letters to related departments or agencies to apply for research funding.

CSE 7	Overall written expression	▪ Can elaborate on abstract topics, using complex sentences, along with a variety of cohesive devices, to construct clear and convincing explanations. ▪ Can collect, analyse, and integrate data into academic writing, providing strong evidence to support his/her viewpoints or refute different opinions. ▪ Can plot complex narratives, using rhetorical devices appropriately so as to enhance liveliness and expressiveness of the language.
	Written description	▪ Can describe data or statistics with accuracy and clarity. ▪ Can summarise the main plots of novels. ▪ Can accurately describe subjective experience and feelings.
	Written narration	▪ Can write complete and standardised news reports about social events. ▪ Can compose stories with complex plots and vivid content.
	Written exposition	▪ Can summarise factual and imaginative texts, highlighting the most important points. ▪ Can write an academic paper in his/her discipline, elaborating on the research process, major challenges, and main findings. ▪ Can properly explain certain social or natural phenomena and analyse tendencies in data and rules of change.
	Written instruction	▪ Can write formal instructions and announcements (e.g. government announcements and court verdicts).
	Written argumentation	▪ Can provide clear viewpoints and/or commentaries on a writer's intent, position, and language style. ▪ Can comment on literary works with clear logic and coherent language. ▪ Can analyse and integrate a variety of data, facts, and views in academic writing. ▪ Can clearly and logically write personal opinion pieces on artistic works (e.g. music and paintings).
	Written interaction	▪ Can write formal and standard conference minutes and formal letters in appropriate language (e.g. invitations to academic conferences).
CSE 6	Overall written expression	▪ Can use various approaches to clarify his/her views on social topical issues or phenomena, with sufficient evidence and logical arguments. ▪ Can write research article abstracts that conform to academic conventions of his/her discipline. ▪ Can produce writing for popular genres (e.g. news reports and book reviews), using appropriate expressions and a generic structure conforming to expected features of style and register.
	Written description	▪ Can describe common goods or products in a relatively accurate and clear manner. ▪ Can briefly describe historical and literary figures in sufficient and salient detail.
	Written narration	▪ Can write science-fiction stories with clear, organised structure and interesting content based on written input or visual prompts. ▪ Can give a generally accurate and complete summary of the plot points of movies or dramas. ▪ Can give clear and organised narrations of well-known Chinese myths, legends, and folklore.
	Written exposition	▪ Can write reports on experiments or surveys with appropriate detail. ▪ Can write a detailed action plan for a project. ▪ Can write an abstract for a research paper, conforming to academic conventions.
	Written instruction	▪ Can write properly-structured business memos in clear language.
	Written argumentation	▪ Can analyse social topical issues on education, entertainment, and people's livelihood and offer specific suggestions. ▪ Can write relatively comprehensive summaries and objective evaluations of articles, books, and movies. ▪ Can analyse research data or case studies in order to support an argument or hypothesis, using accurate language and clear logic. ▪ Can explain the reasons for supporting or opposing a point of view, clearly stating the pros and cons of different viewpoints.

	Written interaction	▪ Can write notices or posters to publicise his/her associations and their activities. ▪ Can write letters of complaint concerning product quality or services with adequate evidence and convincing reasons. ▪ Can write formal letters of congratulation pertaining to successfully-held events.
CSE 5	Overall written expression	▪ Can write short articles with argument and evidence on topics of interest, using a variety of cohesive devices and achieving semantic coherence. ▪ Can write reports related to his/her own field of study (e.g. book reports and survey reports) with complete structure. ▪ Can write for common practical reasons (e.g. letters of gratitude or meeting minutes) with the appropriate use of language.
	Written description	▪ Can describe familiar people or objects with sufficient clarity and explicit expression. ▪ Can describe a familiar scene or setting (e.g. a traditional festival or a celebration) in a detailed manner. ▪ Can describe personal experiences from life and study in relatively detailed content and coherent language.
	Written narration	▪ Can compose short plays with relatively complete plots on familiar themes. ▪ Can narrate events with clear structure and vivid content. ▪ Can provide a relatively coherent narration of his/her experiences.
	Written exposition	▪ Can describe data presented in a graph, table, or chart in a relatively accurate and complete manner. ▪ Can clearly explain the process of campus events (e.g. welcome parties and society meetings). ▪ Can write a relatively detailed introduction to a familiar product, highlighting its main features. ▪ Can clearly explain patterns of distribution shown in a graph and explain how they develop. ▪ Can explain how to deal with daily routines (e.g. returning goods online or making a complaint).
	Written instruction	▪ Can write detailed activity plans for associations or clubs.
	Written argumentation	▪ Can comment on articles or chapters related to his/her study, articulating a clear and convincing viewpoint. ▪ Can discuss social topics of interest, with clear viewpoints and good reasoning. ▪ Can elaborate on his/her views in a relatively clear and orderly way and provide supporting evidence.
	Written interaction	▪ Can write letters of job application, with correct format and complete content, highlighting the salience of his/her qualifications for his/her candidacy. ▪ Can write letters or emails to report campus or social problems to concerned individuals in a clear and complete manner. ▪ Can write letters of sympathy or condolence in appropriate language.
CSE 4	Overall written expression	▪ Can express opinions on topic(s) he/she is familiar with, using some evidence to support his/her viewpoint(s) in a relatively persuasive manner. ▪ Can coherently describe familiar activities (e.g. personal experiences and campus activities), using common rhetorical devices. ▪ Can briefly discuss familiar social and cultural matters (e.g. traditional festivals and customs) through social media (e.g. email and webpages).
	Written description	▪ Can clearly describe familiar places (e.g. hometown and campus). ▪ Can briefly describe changes to familiar people or surroundings. ▪ Can briefly describe favorite movies, including main characters and storyline.
	Written narration	▪ Can write short stories with sufficiently complete storylines based on written or visual input. ▪ Can briefly and coherently narrate important events in his/her life. ▪ Can clearly narrate familiar activities (e.g. class meetings, contests, club activities).
	Written exposition	▪ Can briefly state opinions on common topics. ▪ Can briefly express how he/she feels after taking part in certain social activities. ▪ Can briefly explain familiar but abstract concepts (e.g. friendship or happiness).

	Written instruction	▪ Can briefly describe steps involved in carrying out routine activities. ▪ Can write a clear travel itinerary for outdoor group activities.
	Written argumentation	▪ Can make suggestions on how to solve problems in his/her life or study. ▪ Can clearly explain the pros and cons of a particular action. ▪ Can persuade others to accept his/her points of view.
	Written interaction	▪ Can write letters or emails describing his/her personal information, interests, hobbies, and campus life. ▪ Can write letters or emails to briefly describe familiar places (e.g. scenic spots). ▪ Can briefly outline Chinese culture (e.g. traditional festivals and customs). ▪ Can briefly describe his/her study plans, learning experiences, etc. in different contexts, including social media.
CSE 3	Overall written expression	▪ Can explain in simple terms the causes, processes, and results of events with generally accurate wording. ▪ Can use simple phrases to comment on familiar things and provide reasons, with generally coherent expression. ▪ Can briefly describe his/her daily activities in well-organised written form (e.g. email and WeChat).
	Written description	▪ Can briefly describe recent experiences or mood. ▪ Can briefly describe personal activities (e.g. sports meetings and class meetings). ▪ Can briefly describe favorite places (e.g. school and neighbourhood). ▪ Can briefly describe the salient features of familiar people or objects.
	Written narration	▪ Can write sufficiently complete short stories based on visual input. ▪ Can narrate in simple terms what he/she has experienced, including causes, processes, and results.
	Written exposition	▪ Can clearly explain the process or the procedure of doing something (e.g. planting a tree or borrowing a book from the library). ▪ Can clearly express his/her attitudes towards things around him/her. ▪ Can clearly explain familiar rules and regulations (e.g. school rules or codes of conduct). ▪ Can briefly explain the arrangement of campus events (e.g. school sports meets or class meetings).
	Written instruction	▪ Can clearly explain simple directions (e.g. how to go to the library or the canteen). ▪ Can write simple instructions for common activities based on picture prompts (e.g. how to grow potted plants or how to assemble models).
	Written argumentation	▪ Can briefly comment on something that has just happened around him/her. ▪ Can briefly comment on familiar public figures. ▪ Can briefly comment on some of his/her behaviour.
	Written interaction	▪ Can write letters to familiar people to share mutual experiences. ▪ Can write emails to classmates, detailing specific information (e.g. class schedules or activity instructions). ▪ Can write letters to friends to describe campus and/or family life in a simple way.
CSE 2	Overall written expression	▪ Can describe in simple terms the main features of people or familiar objects in response to prompt(s) (e.g. words and examples). ▪ Can write short stories based on picture(s), using simple words and generating a sufficiently completed storyline. ▪ Can correctly use capitalisation and common punctuation marks.
	Written description	▪ Can describe dreams and wishes with simple short sentences. ▪ Can describe the salient features of familiar people, objects, or places with simple words or sentences. ▪ Can briefly describe family life. ▪ Can briefly describe features of weather or climate. ▪ Can clearly describe personal means of everyday transportation.
	Written narration	▪ Can complete short stories based on visual input. ▪ Can write simple sentences to narrate his/her daily activities. ▪ Can write simple sentences to describe what he/she plans to do.

	Written exposition	▪ Can clearly list his/her daily activities. ▪ Can write simple words and sentences to explain why he/she enjoys (doing) something. ▪ Can write simple words and sentences to state his/her opinions of and attitudes towards certain behaviours (e.g. those pertaining to approval or disapproval).
	Written instruction	▪ Can write down simple game or activity instructions given by a teacher.
	Written argumentation	▪ Can use simple words and sentences to comment on familiar people, events, or objects. ▪ Can write simple sentences to persuade others to do something or dissuade them from doing something. ▪ Can briefly list the reasons why one likes or dislikes some everyday phenomena (e.g. weather and seasons).
	Written interaction	▪ Can write simple notices for class activities. ▪ Can write simple birthday or holiday greeting cards. ▪ Can write simple lost-and-found notices.
CSE 1	Overall written expression	▪ Can correctly copy words and phrases. ▪ Can describe pictures (e.g. animals and foods) using simple words and/or phrases. ▪ Can narrate in simple terms his/her everyday activities based on examples.
	Written description	▪ Can briefly and accurately describe pictures with words or phrases. ▪ Can briefly describe the features of common animals with words or phrases. ▪ Can briefly describe likes and dislikes. ▪ Can label the names of familiar people, places, or objects in pictures.
	Written narration	▪ Can describe in simple terms what he/she is doing. ▪ Can write simple words and sentences to describe familiar daily activities. ▪ Can briefly describe what he/she often does based on written or visual input.
	Written exposition	▪ Can write simple words and phrases to describe his/her personal circumstances. ▪ Can list his/her activities scheduled for a particular day. ▪ Can write simple words and sentences to describe modes of transportation for a return journey.
	Written interaction	▪ Can write simple greetings based on examples. ▪ Can correctly write personal information (e.g. name and age).

Appendix B: Schedule of activities for standard-setting panels

Listening

Day 1

9:30 – 10:00: Introduction to the seminar scope and goals / self-introductions
10:00 – 10:30: Overview of CSE: principals, development methodology, structure
10:30 – 10:45: Break
10:45 – 11:30: Refresher: activities with key descriptors for listening
11:30 – 12:30: Discussion: reflections on the activities in the self-study booklet
12:30 – 14:00: Lunch
14:00 – 15:00: Overview of the tests to be used in the project
15:00 – 16:30: Standard setting with listening items: explanation and practice
16:30 – 16:45: Break
16:45 – 17:45: Final discussion (agree on a list of descriptors, and criterial CSE features)

Day 2

9:00 – 9:45: Review: descriptors, list of criterial CSE features, and Aptis test
9:45 – 11:00: Aptis Listening (BASKET METHOD),
11:00 – 11:30: Break
11:30 – 12:30: Discussion, presentation of item difficulty data,
(Panellists can make adjustments, or re-listen to particular items by choice)
12:30 – 14:00: Lunch
14:00 – 15:30: Aptis Listening (MODIFIED ANGOFF METHOD), Round 1
15:30 – 16:00: Break
16:00 – 16:30: Discussion, presentation of item difficulty data
16:30 – 17:45: Aptis Listening (MODIFIED ANGOFF METHOD), Round 2
17:45 – 18:00: Housekeeping and review plans for final day

Day 3

Day 3 will be used to analyse and review results by the working group

Day 4

9:00 – 9:45: Discussion and review of criterial CSE features and IELTS test
9:45 – 11:00: IELTS Listening (BASKET METHOD),
11:00 – 11:30: Break
11:30 – 12:30: Discussion, presentation of item difficulty data,
(Panellists can make adjustments, or re-listen to particular items by choice)
12:30– 14:00: Lunch
14:00 – 15:30: IELTS Listening (MODIFIED ANGOFF METHOD), Round 1
15:30 – 16:00: Break
16:00 – 16:30: Discussion, presentation of item difficulty data

Reading

Day 1

09:00 Introduction to the workshop scope and goals/self-introductions
09:30 Overview of the CSE: principles, development methodology and structure
10:00 Practice with CSE descriptors based on tasks in the preparation booklet
(break 10:30–10:45)
12:00 Lunch
13:30 Practice with CSE descriptors (cont.)
14:00 Overview of the tests to be used in the project
14:30 Standard setting with reading items: explanation and practice
(break 15:00–15:15)
16:15 Final discussion (confirm a common interpretation of CSE descriptors)
17:15 Finish

Day 2

09:00 Review of distinguishing CSE features, and Aptis test
09:45 Aptis Reading (BASKET METHOD)
(break 11:00–11:15)
11:15 Presentation of item difficulty data, reflection and adjustments
12:00 Lunch
13:30 Aptis Reading (MODIFIED ANGOFF METHOD), Round 1
(break 15:00–15:15)
15:30 Discussion, presentation of item difficulty data
16:00 Aptis Reading (MODIFIED ANGOFF METHOD), Round 2
17:15 Housekeeping and review
17:30 Finish

Day 3

09:00 Review of distinguishing CSE features, and IELTS test
09:45 IELTS Reading (BASKET METHOD)
(break 11:00–11:15)
11:15 Presentation of item difficulty data, reflection and adjustments
12:00 Lunch
13:30 IELTS Reading (MODIFIED ANGOFF METHOD), Round 1
(break 15:00–15:15)
15:30 Discussion, presentation of item difficulty data
16:00 IELTS Reading (MODIFIED ANGOFF METHOD), Round 2
17:15 Feedback questionnaire on standard setting, housekeeping and review
17:45 Finish

Speaking and Writing

Day 1

9:00 Introduction to the workshop scope and goals /self-introductions
9:30 Speaking Refresher: activities with key descriptors for Speaking ability
10:30 Coffee Break
11:00 Finalise Description of Speaking
12:00 Lunch
13:30 Writing Refresher: activities with key descriptors for Writing ability
14:30 Finalise Description of Writing
15:30 Coffee Break
16:00 Overview of the tests to be used in the project
17:00 Standard Setting with Speaking items and Writing: explanation and practice
17:30 Housekeeping and plan for Day 2
18:00 Finish

Day 2

8:30 Discussion and review of distinguishing features of CSE Speaking
9:00 Round 1 judgements of Aptis Speaking performances (including 15 mins break)
12:30 Lunch
14:00 Discussion, presentation of panellists' judgements for common performances
14:30 Round 2 judgements of Aptis speaking performances (including 15 mins break)
17:45 Housekeeping and review plans for final day
18:00 Finish

Day 3

8:30 Discussion and review of distinguishing CSE features and IELTS Speaking test
9:00 Round1 judgements of IELTS speaking performance samples (including 15 mins break)
12:30 Lunch
14:00 Discussion, presentation of panellists' judgements for common items
14:30 Round 2 judgements of IELTS speaking performance samples (including 15 mins break)
17:45 Housekeeping and review plans for final day
18:00 Finish

Day 4

8:30 Discussion and review of distinguishing CSE Writing features and Aptis Writing test
9:00 Round 1 judgements of Aptis Writing performance samples (including 15 mins break)
12:30 Lunch
14:00 Discussion, presentation of panellists' judgements for common items
14:30 Round 2 judgements of Aptis Writing performance samples (including 15 mins break)
17:45 Housekeeping and review plans for final day
18:00 Finish

Day 5

8:30 Discussion and review of distinguishing CSE Writing features and IELTS Writing test
9:00 Round 1 judgements of IELTS Writing performance samples (including 15 mins break)
12:30 Lunch
14:00 Discussion, presentation of panellists' judgements for common items
14:30 Round 2 judgements of IELTS Writing performance samples (including 15 mins break)
17:45 Housekeeping and review plans for final day
18:00 Finish

Appendix C: Completed construct definition templates

IELTS listening

Categories Listening	Section 1, Listening 1	(Listening 1) Item 1	(Listening 1) Item 2	(Listening 1) Item 3	(Listening 1) Item 4	(Listening 1) Item 5
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus						
Task Level (CEFR)	A2					
Response format	Gap fill					
Items per task	7					
Cognitive processing - kind of information targeted	Careful local					
Cognitive processing - levels of listening	Meaning construction					
Content knowledge	1 (General)					
Cultural specificity	2					
Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text
Pattern	Dialogue					
No of speakers	2					
Domain	Public					
Speed						
Discourse mode	Only for monologues					
Nature of information	Only concrete					
Topic	Shopping and obtaining services					
Text genre	Telephone conversations					
Presentation	Verbal (aural)					
Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response
Key information		across sentences	across sentences	Within sentences	Within sentences	across sentences
Operation		Specific information	Specific information	Specific information	Specific information	Specific information
Question presentation		Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)
Option Presentation		NA	NA	NA	NA	NA
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale	3.7.2					
Description	N/A					
Narration	N/A					
Exposition	3.4.3, 2.4.1					
Instruction	N/A					
Argumentation	N/A					
Interaction	2.1.1, 2.1.2; 2.1.3					

Categories Listening	(Listening 1) Item 6	(Listening 1) Item 7	Section 1, Listening 2	(Listening 2) Item 8	(Listening 2) Item 9	(Listening 2) Item 10
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus						
Task Level (CEFR)			B1			
Response format			MCQ			
Items per task			3			
Cognitive processing - kind of information targeted			Careful global			
Cognitive processing - levels of listening			**Meaning construction**			
Content knowledge			1 (General)			
Cultural specificity			2			
Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text
Pattern			Dialogue			
No of speakers			2			
Domain			Public			
Speed						
Discourse mode			Only for monologues			
Nature of information			Mostly concrete			
Topic			Shopping and obtaining services			
Text genre			Telephone conversations			
Presentation			Verbal (aural)			
Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response
Key information	across sentences	across sentences		across sentences	across sentences	across sentences
Operation	Specific information	Specific information		Specific information	Specific information	Specific information
Question presentation	Verbal (written)	Verbal (written)		Verbal (written)	Verbal (written)	Verbal (written)
Option Presentation	NA	NA		Verbal (written)	Verbal (written)	Verbal (written)
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale			4.7.2			
Description			N/A			
Narration			N/A			
Exposition			3.4.1			
Instruction			N/A			
Argumentation			N/A			
Interaction			4.1.3			

Categories Listening	Section 2, Listening 1	(Section 2, Listening 1) Item 11	(Section 2, Listening 1) Item 12	(Section 2, Listening 1) Item 13	(Section 2, Listening 1) Item 14
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus					
Task Level (CEFR)	B1				
Response format	MCQ				
Items per task	6				
Cognitive processing - kind of information targeted	Careful global				
Cognitive processing - levels of listening	Meaning construction				
Content knowledge	1 (General)				
Cultural specificity	2				
Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text
Pattern	Monologue				
No of speakers	2				
Domain	Public				
Speed					
Discourse mode	Only for monologues				
Nature of information	Mostly concrete				
Topic	Descriptions of buildings				
Text genre	Reviews on TV and radio (restaurants, books, movies, etc.)				
Presentation	Verbal (aural)				
Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response
Key information		across paragraphs	across paragraphs	Within sentences	across sentences
Operation		Specific information	Specific information	Specific information	Specific information
Question presentation		Verbal (written)	Verbal (written)	Verbal (written)	Non-verbal (illustrations, graphs, etc)
Option Presentation		Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale	3.7.2				
Description	4.6.1				
Narration	4.5.6				
Exposition	3.4.1				
Instruction	4.3.1				
Argumentation	N/A				
Interaction	N/A				

Categories Listening	(Section 2, Listening 1) Item 15	(Section 2, Listening 1) Item 16	Section 2, Listening 2	(Section 2, Listening 2) Item 17	(Section 2, Listening 2) Item 18	(Section 2, Listening 2) Item 19	(Section 2, Listening 2) Item 20
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus							
Task Level (CEFR)			B1				
Response format			Gap fill				
Items per task			4				
Cognitive processing - kind of information targeted			Careful local				
Cognitive processing - levels of listening			Meaning construction				
Content knowledge			1 (General)				
Cultural specificity			2				
Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text
Pattern			Monologue				
No of speakers			1				
Domain			Public				
Speed							
Discourse mode			Descriptive				
Nature of information			Mostly concrete				
Topic			Descriptions of buildings				
Text genre			Reviews on TV and radio (restaurants, books, movies, etc.)				
Presentation			Verbal (aural)				
Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response
Key information	across sentences	across sentences		Within sentences	Within sentences	Within sentences	Within sentences
Operation	Specific information	Specific information		Specific information	Specific information	Specific information	Specific information
Question presentation	Non-verbal (illustrations, graphs, etc)	Non-verbal (illustrations, graphs, etc)		Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)
Option Presentation	Verbal (written)	Verbal (written)		NA	NA	NA	NA
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale			4.7.2				
Description			4.6.2; 4.6.1				
Narration			N/A				
Exposition			N/A				
Instruction			N/A				
Argumentation			N/A				
Interaction			N/A				

Categories Listening	Section 3, Listening 1	(Section 3, Listening 1) Item 21	(Section 3, Listening 1) Item 22	(Section 3, Listening 1) Item 23	(Section 3, Listening 1) Item 24	(Section 3, Listening 1) Item 25
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus						
Task Level (CEFR)	B1					
Response format	Gap fill					
Items per task	5					
Cognitive processing - kind of information targeted	Careful global					
Cognitive processing - levels of listening	Meaning construction					
Content knowledge	2					
Cultural specificity	2					
Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text
Pattern	Dialogue					
No of speakers	3					
Domain	Educational					
Speed						
Discourse mode	Only for monologues					
Nature of information	Mostly concrete					
Topic	Education — college life					
Text genre	Interpersonal dialogues and conversations					
Presentation	Verbal (aural)					
Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response
Key information		across sentences	across sentences	across sentences	Within sentences	Within sentences
Operation		Specific information	Specific information	Specific information	Specific information	Specific information
Question presentation		Verbal (written)	Verbal (written)	Verbal (written)	Non-verbal (illustrations, graphs, etc)	Non-verbal (illustrations, graphs, etc)
Option Presentation		NA	NA	NA	NA	NA
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale	3.7.2					
Description	N/A					
Narration	N/A					
Exposition	N/A					
Instruction	N/A					
Argumentation	N/A					
Interaction	4.1.1, 4.1.3					

Categories Listening	Section 3, Listening 2	(Section 3, Listening 2) Item 26	(Section 3, Listening 2) Item 27	(Section 3, Listening 2) Item 28	(Section 3, Listening 2) Item 29	(Section 3, Listening 2) Item 30
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus						
Task Level (CEFR)	B1					
Response format	MCQ					
Items per task	5					
Cognitive processing - kind of information targeted	Careful global					
Cognitive processing - levels of listening	Meaning construction					
Content knowledge	1 (General)					
Cultural specificity	2					
Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text
Pattern	Dialogue					
No of speakers	3					
Domain	Educational					
Speed						
Discourse mode	Only for monologues					
Nature of information	Mostly concrete					
Topic	Education — college life					
Text genre	Interpersonal dialogues and conversations					
Presentation	Verbal (aural)					
Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response
Key information		across sentences	across sentences	Within sentences	Within sentences	Within sentences
Operation		Specific information	Specific information	Specific information	Specific information	Specific information
Question presentation		Verbal (written)	Verbal (written)	Non-verbal (illustrations, graphs, etc)	Non-verbal (illustrations, graphs, etc)	Non-verbal (illustrations, graphs, etc)
Option Presentation		Verbal (written)	Non-verbal (illustrations, graphs, etc)	NA	NA	NA
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale	4.7.2					
Description	N/A					
Narration	N/A					
Exposition	3.4.1					
Instruction	N/A					
Argumentation	N/A					
Interaction	4.1.3					

Categories Listening	Section 4, Listening 1	(Section 4, Listening 1) Item 31	(Section 4, Listening 1) Item 32	(Section 4, Listening 1) Item 33	(Section 4, Listening 1) Item 34
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus					
Task Level (CEFR)	B2				
Response format	Gap fill				
Items per task	10				
Cognitive processing - kind of information targeted	Careful global				
Cognitive processing - levels of listening	Meaning construction				
Content knowledge	2				
Cultural specificity	2				
Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text
Pattern	Monologue				
No of speakers	1				
Domain	Educational				
Speed					
Discourse mode	Instructive				
Nature of information	Fairly abstract				
Topic	Education — college life				
Text genre	Public speeches, lectures, presentations				
Presentation	Verbal (aural)				
Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response
Key information		Within sentences	Within sentences	across sentences	Within sentences
Operation		Specific information	Specific information	Specific information	Specific information
Question presentation		Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)
Option Presentation		NA	NA	NA	NA
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale	6.7.2, 8.7.2				
Description	N/A				
Narration	N/A				
Exposition	5.4.3				
Instruction	5.3.2				
Argumentation	7.2.1				
Interaction	N/A				

Categories Listening	(Section 4, Listening 1) Item 35	(Section 4, Listening 1) Item 36	(Section 4, Listening 1) Item 37	(Section 4, Listening 1) Item38	(Section 4, Listening 1) Item 39	(Section 4, Listening 1) Item 40
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus						
Task Level (CEFR)						
Response format						
Items per task						
Cognitive processing - kind of information targeted						
Cognitive processing - levels of listening						
Content knowledge						
Cultural specificity						
Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text
Pattern						
No of speakers						
Domain						
Speed						
Discourse mode						
Nature of information						
Topic						
Text genre						
Presentation						
Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response
Key information	Within sentences	Within sentences	Within sentences	Within sentences	Within sentences	Within sentences
Operation	Specific information	Specific information	Specific information	Specific information	Specific information	Specific information
Question presentation	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)
Option Presentation	NA	NA	NA	NA	NA	NA
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale						
Description						
Narration						
Exposition						
Instruction						
Argumentation						
Interaction						

IELTS reading

Categories Reading	Task 1	(Task 1) Item 1	(Task 1) Item 2	(Task 1) Item 3	(Task 1) Item 4	(Task 1) Item 5	(Task 1) Item 6
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus							
Task Level (CEFR)	B2						
Response format	MCQ						
Items per task							
Cognitive processing 1	Expeditious reading: global						
Cognitive processing 2	Building a mental model						
Content knowledge	2						
Cultural specificity	1 (Neutral)						
Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text
Domain	Public						
Discourse mode	Expository						
Nature of information	Fairly abstract						
Topic	Descriptions of people (appearance, personality)						
Text genre	Magazines						
Presentation	Verbal (written)						
Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response
Key information		across sentences	Within sentences	Within sentences	across sentences	Within sentences	across sentences
Operation		Specific information	Specific information	Specific information	Specific information	Specific information	Specific information
Question presentation		Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)
Option Presentation		Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale	6.7.3						
Description	N/A						
Narration	N/A						
Exposition	6.4.4, 6.4.5						
Instruction	N/A						
Argumentation	6.2.7						
Interaction	N/A						

Categories Reading	(Task 1) Item 7	(Task 1) Item 8	(Task 1) Item 9	(Task 1) Item 10	(Task 1) Item 11	(Task 1) Item 12	(Task 1) Item 13
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus							
Task Level (CEFR)							
Response format							
Items per task							
Cognitive processing 1							
Cognitive processing 2							
Content knowledge							
Cultural specificity							
Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text
Domain							
Discourse mode							
Nature of information							
Topic							
Text genre							
Presentation							
Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response
Key information	across sentences	across sentences	Within sentences	Within sentences	Within sentences	Within sentences	Within sentences
Operation	Specific information	Specific information	Specific information	Specific information	Specific information	Specific information	Specific information
Question presentation	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)
Option Presentation	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale							
Description							
Narration							
Exposition							
Instruction							
Argumentation							
Interaction							

Categories Reading	Task 2	(Task 2) Item 14	(Task 2) Item 15	(Task 2) Item 16	(Task 2) Item 17	(Task 2) Item 18	(Task 2) Item 19
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus							
Task Level (CEFR)	B2						
Response format	MCQ						
Items per task							
Cognitive processing 1	Careful reading: local						
Cognitive processing 2	Establishing propositional meaning (cl./sent. level)						
Content knowledge	3						
Cultural specificity	2						
Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text
Domain	Public						
Discourse mode	Expository						
Nature of information	Mostly concrete						
Topic	Plants, animals, nature						
Text genre	Magazines						
Presentation	Verbal (written)						
Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response
Key information		Within sentences	Within sentences	across sentences	Within sentences	Within sentences	across sentences
Operation		Specific information	Specific information	Specific information	Specific information	Specific information	Specific information
Question presentation		Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)
Option Presentation		Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale	6.7.3						
Description	N/A						
Narration	N/A						
Exposition	7.4.5, 7.4.6, 6.4.4						
Instruction	N/A						
Argumentation	7.2.6						
Interaction	N/A						

Categories Reading	(Task 2) Item 20	(Task 2) Item 21	(Task 2) Item 22	(Task 2) Item 23	(Task 2) Item 24	(Task 2) Item 25	(Task 2) Item 26
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus							
Task Level (CEFR)							
Response format							
Items per task							
Cognitive processing 1							
Cognitive processing 2							
Content knowledge							
Cultural specificity							
Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text
Domain							
Discourse mode							
Nature of information							
Topic							
Text genre							
Presentation							
Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response
Key information	Within sentences	Within sentences	Within sentences	Within sentences	Within sentences	Within sentences	across sentences
Operation	Specific information	Specific information	Specific information	Specific information	Specific information	Specific information	Specific information
Question presentation	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)
Option Presentation	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale							
Description							
Narration							
Exposition							
Instruction							
Argumentation							
Interaction							

Categories Reading	Task 3	(Task 3) Item 27	(Task 3) Item 28	(Task 3) Item 29	(Task 3) Item 30
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus					
Task Level (CEFR)	C1				
Response format	Matching headings to text				
Items per task					
Cognitive processing 1	Careful reading: global				
Cognitive processing 2	Creating a text level representation (disc. structure)				
Content knowledge	2				
Cultural specificity	1 (Neutral)				
Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text
Domain	Public				
Discourse mode	Expository				
Nature of information	Mostly abstract				
Topic	Science and technology				
Text genre	Magazines				
Presentation	Verbal (written)				
Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response
Key information		across paragraphs	across paragraphs	across paragraphs	across paragraphs
Operation		Main idea / conclusions	Main idea / conclusions	Main idea / conclusions	Main idea / conclusions
Question presentation		Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)
Option Presentation		Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale	8.7.1				
Description	N/A				
Narration	N/A				
Exposition	7.4.6, 7.4.5				
Instruction	N/A				
Argumentation	8.2.3, 7.2.3				
Interaction	N/A				

Categories Reading	(Task 3) Item 31	(Task 3) Item 32	(Task 3) Item 33	(Task 3) Item 34	(Task 3) Item 35
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus					
Task Level (CEFR)					
Response format					
Items per task					
Cognitive processing 1					
Cognitive processing 2					
Content knowledge					
Cultural specificity					
Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text
Domain					
Discourse mode					
Nature of information					
Topic					
Text genre					
Presentation					
Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response
Key information	across paragraphs	across sentences	Within sentences	across sentences	Within sentences
Operation	Main idea / conclusions	Opinion	Opinion	Opinion	Opinion
Question presentation	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)
Option Presentation	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale					
Description					
Narration					
Exposition					
Instruction					
Argumentation					
Interaction					

Categories Reading	(Task 3) Item 36	(Task 3) Item 37	(Task 3) Item 38	(Task 3) Item 39	(Task 3) Item 40
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus					
Task Level (CEFR)					
Response format					
Items per task					
Cognitive processing 1					
Cognitive processing 2					
Content knowledge					
Cultural specificity					
Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text
Domain					
Discourse mode					
Nature of information					
Topic					
Text genre					
Presentation					
Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response
Key information	across sentences	across sentences	Within sentences	across sentences	Text level representation
Operation	Opinion	Opinion	Opinion	Opinion	Main idea / conclusions
Question presentation	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)
Option Presentation	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale					
Description					
Narration					
Exposition					
Instruction					
Argumentation					
Interaction					

IELTS speaking

Categories Speaking	Task 1	(Task 1) Topic 1	(Task 1) Topic 2	(Task 1) Topic 3	(Task 1) Topic 4	(Task 1) Topic 5	(Task 1) Topic 6
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus							
Task Level (CEFR)	B1						
Response format	Constructed response - spoken						
Planning time (y/n)	No						
Planning time (seconds)	N/A						
Pattern of interaction	Interlocutor-Candidate (I-C)						
Nature of intelocutor input 1	Scripted						
Nature of interlocutor input 2	N/A						
Informational functions (multiple selection possible)	1,2,3,4,5,8,10,11						
Interactional functions (multiple selection possible)	20						
Managing interaction functions (multiple selection possible)	N/A						
Content knowledge	1 (General)						
Cultural specificity	1 (Neutral)						
Features of the input	Features of the input	Features of the input	Features of the input	Features of the input	Features of the input	Features of the input	Features of the input
Description	A sentence to introduce the question topic + four direct questions	A sentence to introduce the question topic + four direct questions	A sentence to introduce the question topic + four direct questions	A sentence to introduce the question topic + four direct questions	A sentence to introduce the question topic + four direct questions	A sentence to introduce the question topic + four direct questions	A sentence to introduce the question topic + four direct questions
Domain	Personal	Personal					
Nature of information	Mostly concrete						
Topic							
Presentation	Verbal (aural)	Verbal (aural)	Verbal (aural)	Verbal (aural)	Verbal (aural)	Verbal (aural)	Verbal (aural)
Features of the Response							
Expected ouput (description)		Short responses (e.g. 1 - 2 'sentences')ON THE SPECIFIC TOPIC OF NAMES 1_PROVIDING PERSONAL INFO 2_EXPLAINING OPINIONS/PREFERENCES 11_EXPRESSING PREFERENCES ... TAPPING INTO THESE BROAD TOPICS CULTURE AND CUSTOMS RELATIONSHIPS AND FAMILY SOCIAL TRENDS	Short responses (e.g. 1 - 2 'sentences') ON THE SPECIFIC TOPIC OF COMPUTERS 1_PROVIDING PERSONAL INFO 2_EXPLAINING OPINIONS/PREFERENCES 11_EXPRESSING PREFERENCES ... TAPPING INTO THESE BROAD TOPICS DAILY LIFE TRAINING AND LEARNING DREAMS AND FUTURE PLANS	Short responses (e.g. 1 - 2 'sentences') ON THE SPECIFIC TOPIC OF HAPPINESS 1_PROVIDING PERSONAL INFO 2_EXPLAINING OPINIONS/PREFERENCES 4_JUSTIFYING OPINIONS 3_ELABORATING ... TAPPING INTO THESE BROAD TOPICS PERSONAL FINANCES HEALTH AND MEDICINE - SOCIAL TOPIC	Short responses (e.g. 1 - 2 'sentences') ON THE SPECIFIC TOPIC OF BIRDS 2_EXPLAINING OPINIONS/PREFERENCES 4_JUSTIFYING OPINIONS 11_EXPRESSING PREFERENCES 3_ELABORATING ... TAPPING INTO THESE BROAD TOPICS DESCRIPTIONS OF PLACES PLANTS, ANIMALS, NATURE CUSTOMS AND CULTURE ENVIRONMENTAL ISSUES	Short responses (e.g. 1 - 2 'sentences') ON THE SPECIFIC TOPIC OF PHOTOS 1_PROVIDING PERSONAL INFO 2_EXPLAINING OPINIONS/PREFERENCES 4_JUSTIFYING OPINIONS 5_COMPARING 11_EXPRESSING PREFERENCES 8_DESCRIBING 3_ELABORATING ... TAPPING INTO THESE BROAD TOPICS Shopping and obtaining services Descriptions Social trends Travel and Tourism	Short responses (e.g. 1 - 2 'sentences') ON THE TOPIC OF FOOD 2_EXPLAINING OPINIONS/PREFERENCES 4_JUSTIFYING OPINIONS 5_COMPARING 8_DESCRIBING 11_EXPRESSING PREFERENCES ... TAPPING INTO THESE BROAD TOPICS Food and drink Social trends Health and Medicine – social topic Daily life
Form		Spoken	Spoken	Spoken	Spoken	Spoken	Spoken
Expected output (time in minutes)		1-2 mins	1-2 mins	1-2 mins	1-2 mins	1-2 mins	1-2 mins
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale	4.7.2, 4.7.3						
Description	4.6.2						
Narration	N/A						
Exposition	N/A						
Instruction	N/A						
Argumentation	3.2.3, 4.2.1						
Interaction	N/A						

Categories Speaking	Task 2	(Task 2) Item 1	Task 3	(Task 3) Topic 1	(Task 3) Topic 2 (if used)
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus					
Task Level (CEFR)	B1		B2		
Response format	Constructed response - spoken		Constructed response - spoken		
Planning time (y/n)	Yes		No		
Planning time (seconds)	60		N/A		
Pattern of interaction	Interlocutor-Candidate (I-C)		Interlocutor-Candidate (I-C)		
Nature of intelocutor input 1	Scripted		guided		
Nature of interlocutor input 2	Scripted		guided		
Informational functions (multiple selection possible)	1,2,3,4,6,7,8,9,11		2,3,4,5,6,8,9,10,11		
Interactional functions (multiple selection possible)	15,19,20		13,14,15,17,19,20		
Managing interaction functions (multiple selection possible)	N/A		22,24		
Content knowledge	1 (General)		1 (General)		
Cultural specificity	1 (Neutral)		2		
Features of the input	Features of the input	Features of the input	Features of the input	Features of the input	Features of the input
Description	examiner script introducing the topic and giving task instructions + a prompt card (rubric and 3 bullet points with suggestions on what to talk about)	introducing the topic and giving task instructions + a prompt card (rubric and 3 bullet points with suggestions on what to talk about)	introducing the task (fully scripted) + 3 themes with 3 bullet points each containing suggestions for examiners on how to develop the discussion. The themes get progressively more abstract and points for discussion more complex	a theme (sub-topic) with 3 bullet points containing suggestions for discussion	a theme (sub-topic) with 3 bullet points containing suggestions for discussion
Domain	Personal		Public	Public	Public
Nature of information	Mostly concrete		Fairly abstract	Fairly abstract	Fairly abstract
Topic	Travel and tourism		Travel and tourism	Travel and tourism	Social trends
Presentation	Written & aural	Written & aural	Verbal (aural)	Verbal (aural)	Verbal (aural)
Features of the Response					
Expected ouput (description)		Extended turn (mainly monologic) followed by real-time processing of questions from the interlocutor and real-time responses to these.		two-way real-time discussion with the interlocutor	two-way real-time discussion with the interlocutor
Form		Spoken		Spoken	Spoken
Expected output (time in minutes)		1 to 2 min		4-5 min	4-5 min
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale	5.7.1, 5.7.3		7.7.2, 7.7.3		
Description	4.6.5, 5.6.1		6.6.2		
Narration	N/A		N/A		
Exposition	5.4.5, 5.4.6		6.4.2, 6.4.3, 6.4.4		
Instruction	N/A		N/A		
Argumentation	5.2.4		6.2.3, 6.2.4, 6.2.5		
Interaction	5.1.8, 4.1.9		6.1.2, 7.1.1, 7.1.2		

IELTS writing

Categories Writing	Task 1	(Task 1) Item 1	Task 2	(Task 2) Item 1
Features of the Task	**Features of the Task**	**Features of the Task**	**Features of the Task**	**Features of the Task**
Skill focus				
Task Level (CEFR)	B1		B2	
Response format	Contructed response - written		Contructed response - written	
Planning time (y/n)	Yes		Yes	
Planning time (seconds)	unspecified - included in the writing time (20 min)		unspecified - included in the writing time (40 min)	
Functions (BC EAQUALS list)	6,31,34		19,21,22,23,28,31,32,33,34,35,36,38	
Content knowledge	2		2	
Cultural specificity	2		2	
Features of the input	**Features of the input**	**Features of the input**	**Features of the input**	**Features of the input**
Description	4 sentence rubric + graphic input		4-sentence rubric + two-part prompt	
Domain	Occupational		Educational	
Nature of information	Mostly concrete		Fairly abstract	
Topic	Work and job related		Education — school life	
Presentation	Non-verbal (illustrations, graphs, etc)		Verbal (written)	
Features of the Response	**Features of the Response**	**Features of the Response**	**Features of the Response**	**Features of the Response**
Expected ouput (description)		a brief academic-style report describing a process		an academic-style, opinon-based, argumentative essay
Expected output (number of words)		150 words minimum		250 words minimum
Form of written output		Written		Written
Discourse mode		Descriptive		Argumentative
Intended genre		Other		Other
CSE Descriptors	**CSE Descriptors**	**CSE Descriptors**	**CSE Descriptors**	**CSE Descriptors**
Overall scale	5.7.2		6.7.3	
Description	N/A		5.6.3	
Narration	N/A		5.5.1	
Exposition	5.4.5, 3.4.4		4.4.1, 4.4.3	
Instruction	3.3.1		N/A	
Argumentation	N/A		6.2.1,6.2.4, 5.2.1	
Interaction	N/A		N/A	

Aptis Listening

Categories Listening	Task 1	(Task 1) Item 1	Task 2	(Task 2) Item 1	Task 3	(Task 3) Item 1
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus	Picking out key word		Picking out key word		Picking out key word	
Task Level (CEFR)	A1		A2		A1	
Response format						
Items per task						
Cognitive processing - kind of infomration targeted	Careful local		Careful local		Careful local	
Cognitive processing - levels of listening	Lexical search		Lexical search		Lexical search	
Content knowledge	1 (General)		1 (General)		1 (General)	
Cultural specificity	3		1 (Neutral)		1 (Neutral)	
Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text
Pattern	Monologue		Monologue		Monologue	
No of speakers	1		1		1	
Domain	Public		Public		Personal	
Speed						
Discourse mode	Expository		Expository		Expository	
Nature of information	Only concrete		Only concrete		Only concrete	
Topic	Daily life		Daily life		Daily life	
Text genre	Telephone — voicemail and answering machine messages		Telephone — pre-recorded information services		Telephone — voicemail and answering machine messages	
Presentation						
Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response
Key information		across sentences		across sentences		across sentences
Operation		Specific information		Specific information		Specific information
Question presentation						
Option Presentation						
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale	2.7.1		2.7.1		2.7.1	
Description	N/A		N/A		N/A	
Narration	N/A		N/A		N/A	
Exposition	2.4.1		2.4.1		N/A	
Instruction	N/A		N/A		N/A	
Argumentation	N/A		N/A		N/A	
Interaction	N/A		N/A		N/A	

Categories Listening	Task 4	(Task 4) Item 1	Task 5	(Task 5) Item 1	Task 6	(Task 6) Item 1
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus	Picking out key word		Picking out key word			
Task Level (CEFR)	A1		A1		A2	
Response format						
Items per task						
Cognitive processing - kind of infomration targeted	Careful local		Careful local		Careful global	
Cognitive processing - levels of listening	**Lexical search**		**Lexical search**		**Lexical search**	
Content knowledge	1 (General)		1 (General)		1 (General)	
Cultural specificity	1 (Neutral)		1 (Neutral)		1 (Neutral)	
Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text
Pattern	Monologue		Monologue		Dialogue	
No of speakers	1		1		2	
Domain	Occupational		Personal		Personal	
Speed						
Discourse mode	Expository		Expository		Only for monologues	
Nature of information	Only concrete		Only concrete		Mostly concrete	
Topic	Work and job related		Leisure and entertainment		Travel and tourism	
Text genre	Telephone — voicemail and answering machine messages		Telephone conversations		Interpersonal dialogues and conversations	
Presentation						
Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response
Key information		across sentences		across sentences		across sentences
Operation		Specific information		Specific information		Specific information
Question presentation						
Option Presentation						
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale	2.7.1		2.7.1		2.7.1, 3.7.2	
Description	N/A		N/A		N/A	
Narration	N/A		N/A		3.5.2	
Exposition	2.4.1		2.4.1		N/A	
Instruction	N/A		N/A		N/A	
Argumentation	N/A		N/A		N/A	
Interaction	N/A		N/A		N/A	

Categories Listening	Task 7	(Task 7) Item 1	Task 8	(Task 8) Item 1	Task 9	(Task 9) Item 1
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus						
Task Level (CEFR)	A2		A2		A2	
Response format						
Items per task						
Cognitive processing - kind of infomration targeted	Careful local		Careful local		Careful local	
Cognitive processing - levels of listening	Lexical search		Lexical search		Lexical search	
Content knowledge	1 (General)		1 (General)		1 (General)	
Cultural specificity	1 (Neutral)		1 (Neutral)		1 (Neutral)	
Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text
Pattern	Monologue		Monologue		Monologue	
No of speakers	1		1		1	
Domain	Public		Personal		Personal	
Speed						
Discourse mode	Expository		Expository		Expository	
Nature of information	Only concrete		Mostly concrete		Only concrete	
Topic	Education — college life		Leisure and entertainment		Food and drink	
Text genre	Public speeches, lectures, presentations		Other		Interpersonal dialogues and conversations	
Presentation						
Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response
Key information		across sentences		across sentences		across sentences
Operation		Specific information		Specific information		Specific information
Question presentation						
Option Presentation						
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale	2.7.1		2.7.1, 3.7.2		2.7.1, 3.7.2	
Description	N/A		N/A		N/A	
Narration	N/A		N/A		N/A	
Exposition	3.4.2		3.4.2		N/A	
Instruction	N/A		N/A		N/A	
Argumentation	N/A		N/A		N/A	
Interaction	N/A		N/A		N/A	

Categories Listening	Task 10	(Task 10) Item 1	Task 11	(Task 11) Item 1	Task 12	(Task 12) Item 1
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus			Idetifying specific information			
Task Level (CEFR)	A2		B1		B1	
Response format						
Items per task						
Cognitive processing - kind of infomration targeted	Careful local		Careful local		Careful local	
Cognitive processing - levels of listening	Lexical search		Meaning construction		Meaning construction	
Content knowledge	1 (General)		3		5 (Specific	
Cultural specificity	1 (Neutral)		2		2	
Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text
Pattern	Monologue		Monologue		Monologue	
No of speakers	1		1		1	
Domain	Occupational		Educational		Educational	
Speed						
Discourse mode	Expository		Expository		Expository	
Nature of information	Only concrete		Mostly concrete		Only concrete	
Topic	Work and job related		Education — training and learning		Work and job related	
Text genre	Telephone — voicemail and answering machine messages		Interpersonal dialogues and conversations		Public announcements and instructions	
Presentation						
Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response
Key information		across sentences		across sentences		across sentences
Operation		Specific information		Specific information		Specific information
Question presentation						
Option Presentation						
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale	2.7.1, 3.7.2		4.7.2		4.7.2	
Description	N/A		N/A		N/A	
Narration	N/A		N/A		N/A	
Exposition	2.4.1		N/A		N/A	
Instruction	N/A		N/A		4.3.2	
Argumentation	N/A		N/A		N/A	
Interaction	3.1.3		4.1.2		N/A	

Categories Listening	Task 13	(Task 13) Item 1	Task 14	(Task 14) Item 1	(Task 14) Item 2	(Task 14) Item 3
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus						
Task Level (CEFR)	B1		B1			
Response format						
Items per task						
Cognitive processing - kind of infomration targeted	Careful local		Expeditious global			
Cognitive processing - levels of listening	Meaning construction		Meaning construction			
Content knowledge	2		1 (General)			
Cultural specificity	1 (Neutral)		1 (Neutral)			
Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text
Pattern	Monologue		Monologue			
No of speakers	1		4			
Domain	Personal		Personal			
Speed						
Discourse mode	Expository		Expository			
Nature of information			Only concrete			
Topic	Travel and tourism		Dreams and future plans			
Text genre	Telephone conversations		Interviews (both live and on broadcast media)			
Presentation						
Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response
Key information		across sentences		across sentences	across sentences	across sentences
Operation		Specific information		Gist	Gist	Gist
Question presentation						
Option Presentation						
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale	4.7.2		4.7.2			
Description	4.6.2		N/A			
Narration	N/A		N/A			
Exposition	N/A		N/A			
Instruction	N/A		N/A			
Argumentation	N/A		N/A			
Interaction	N/A		N/A			

Categories Listening	(Task 14) Item 4	Task 15	(Task 15) Item 1	(Task 15) Item 2	Task 16	(Task 16) Item 1
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus		Identifying attiude and opinion			Identifying attiude and opinion	
Task Level (CEFR)		B2			B2	
Response format						
Items per task						
Cognitive processing - kind of infomration targeted		Expeditious global			Expeditious global	
Cognitive processing - levels of listening		**Meaning construction**			**Meaning construction**	
Content knowledge		4			5 (Specific	
Cultural specificity		4			3	
Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text
Pattern		Monologue			Monologue	
No of speakers		1			1	
Domain		Public			Public	
Speed						
Discourse mode		Argumentative			Argumentative	
Nature of information		Fairly abstract			Fairly abstract	
Topic		Politics and government			Culture and customs	
Text genre		Public speeches, lectures, presentations			Reviews on TV and radio (restaurants, books, movies, etc.)	
Presentation						
Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response
Key information	across sentences		Text level representation	Text level representation		Text level representation
Operation	Gist		Opinion	Opinion		Opinion
Question presentation						
Option Presentation						
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale		7.7.2.			7.7.2	
Description		N/A			N/A	
Narration		N/A			N/A	
Exposition		N/A			N/A	
Instruction		N/A			N/A	
Argumentation		7.2.4			7.2.4	
Interaction		N/A			N/A	

Categories Listening	(Task 16) Item 2	Task 17	(Task 17) Item 1	(Task 17) Item 2	(Task 17) Item 3	(Task 17) Item 4
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus						
Task Level (CEFR)		B2				
Response format						
Items per task						
Cognitive processing - kind of infomration targeted		Expeditious global				
Cognitive processing - levels of listening		Meaning construction				
Content knowledge		3				
Cultural specificity		2				
Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text
Pattern		Dialogue				
No of speakers		1				
Domain		Personal				
Speed						
Discourse mode		Only for monologues				
Nature of information		Fairly abstract				
Topic		Health & medicine -- social topic				
Text genre		Interpersonal dialogues and conversations				
Presentation						
Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response
Key information	Text level representation		across sentences	across sentences	across sentences	across sentences
Operation	Opinion		Opinion	Opinion	Opinion	Opinion
Question presentation						
Option Presentation						
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale		5.7.2.				
Description		N/A				
Narration		N/A				
Exposition		3.4.3				
Instruction		N/A				
Argumentation		N/A				
Interaction		5.1.2				

Aptis Reading

Categories Reading	Task 1	(Task 1) Item 1	(Task 1) Item 2	(Task 1) Item 3	(Task 1) Item 4	(Task 1) Item 5
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus	Meaning at sentence level					
Task Level (CEFR)	A1					
Response format	Multiple choice gap fill					
Items per task	5					
Cognitive processing 1	Careful reading: local					
Cognitive processing 2	Lexical access					
Content knowledge	1 (General)					
Cultural specificity	2					
Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text
Domain	Personal					
Discourse mode	Descriptive					
Nature of information	Only concrete					
Topic	Travel and tourism					
Text genre	Personal letters / e-mail					
Presentation	Verbal (written)					
Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response
Key information		Within sentences	Within sentences	Within sentences	Within sentences	Within sentences
Operation		Specific information	Specific information	Specific information	Specific information	Specific information
Question presentation						
Option Presentation						
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale	2.7.2					
Description	N/A					
Narration	2.5.1, 2.5.4					
Exposition	N/A					
Instruction	N/A					
Argumentation	N/A					
Interaction	2.1.1					

Categories Reading	Task 2	(Task 2) Item 1	(Task 2) Item 2	(Task 2) Item 3	(Task 2) Item 4	(Task 2) Item 5
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus	Cohesion between sentences					
Task Level (CEFR)	A2					
Response format						
Items per task	5					
Cognitive processing 1	Careful reading: global					
Cognitive processing 2	Creating a text level representation (disc. structure)					
Content knowledge	1 (General)					
Cultural specificity	1 (Neutral)					
Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text
Domain	Public					
Discourse mode	Expository					
Nature of information	Only concrete					
Topic	Travel and tourism					
Text genre	Reports and memorandums					
Presentation	Verbal (written)					
Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response
Key information		across sentences	across sentences	across sentences	across sentences	across sentences
Operation		Text structure / connections between the parts	Text structure / connections between the parts	Text structure / connections between the parts	Text structure / connections between the parts	Text structure / connections between the parts
Question presentation						
Option Presentation						
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale	3.7.1					
Description	N/A					
Narration	N/A					
Exposition	3.4.2					
Instruction	N/A					
Argumentation	N/A					
Interaction	N/A					

Categories Reading	Task 3	(Task 3) Item 1	(Task 3) Item 2	(Task 3) Item 3	(Task 3) Item 4	(Task 3) Item 5
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus	Cohesion between sentences					
Task Level (CEFR)	A2					
Response format						
Items per task	5					
Cognitive processing 1	Careful reading: global					
Cognitive processing 2	Creating a text level representation (disc. structure)					
Content knowledge	1 (General)					
Cultural specificity	2					
Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text
Domain	Public					
Discourse mode	Instructive					
Nature of information	Only concrete					
Topic	Work and job related					
Text genre	Instructional materials (handouts, textbooks, etc.)					
Presentation	Verbal (written)					
Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response
Key information		across sentences	across sentences	across sentences	across sentences	across sentences
Operation		Text structure / connections between the parts	Text structure / connections between the parts	Text structure / connections between the parts	Text structure / connections between the parts	Text structure / connections between the parts
Question presentation						
Option Presentation						
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale	3.7.1					
Description	N/A					
Narration	N/A					
Exposition	3.4.2					
Instruction	N/A					
Argumentation	N/A					
Interaction	N/A					

Categories Reading	Task 4	(Task 4) Item 1	(Task 4) Item 2	(Task 4) Item 3	(Task 4) Item 4	(Task 4) Item 5	(Task 4) Item 6	(Task 4) Item 7
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus	Comprehension of short texts							
Task Level (CEFR)	B1							
Response format	MCQ							
Items per task	7							
Cognitive processing 1	Careful reading: global							
Cognitive processing 2	Establishing propositional meaning (cl./sent. level)							
Content knowledge	1 (General)							
Cultural specificity	1 (Neutral)							
Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text
Domain	Educational							
Discourse mode	Argumentative							
Nature of information	Fairly abstract							
Topic	Education — training and learning							
Text genre	Magazines							
Presentation	Verbal (written)							
Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response
Key information		across sentences	across sentences	across sentences	across sentences	across sentences	across sentences	across sentences
Operation		Writer's / speaker's attitude	Writer's / speaker's attitude	Writer's / speaker's attitude	Writer's / speaker's attitude	Writer's / speaker's attitude	Writer's / speaker's attitude	Writer's / speaker's attitude
Question presentation								
Option Presentation								
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale	4.7.2							
Description	N/A							
Narration	N/A							
Exposition	N/A							
Instruction	N/A							
Argumentation	4.2.6. 4.2.5							
Interaction	N/A							

Categories Reading	Task 5	(Task 5) Item 1	(Task 5) Item 2	(Task 5) Item 3	(Task 5) Item 4	(Task 5) Item 5	(Task 5) Item 6	(Task 5) Item 7
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus	Text level comprehension, longer text							
Task Level (CEFR)	B2							
Response format	Matching headings to text							
Items per task	7							
Cognitive processing 1	Expeditious reading: global							
Cognitive processing 2	Building a mental model							
Content knowledge	2							
Cultural specificity	4							
Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text	Features of the Input Text
Domain	Educational							
Discourse mode	Narrative							
Nature of information	Only concrete							
Topic	Weather							
Text genre	Magazines							
Presentation	Verbal (written)							
Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response
Key information		across paragraphs	across paragraphs	across paragraphs	across paragraphs	across paragraphs	across paragraphs	across paragraphs
Operation		Text structure / connections between the parts	Text structure / connections between the parts	Text structure / connections between the parts	Text structure / connections between the parts	Text structure / connections between the parts	Text structure / connections between the parts	Text structure / connections between the parts
Question presentation								
Option Presentation								
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale	5.7.2							
Description	N/A							
Narration	5.5.1							
Exposition	6.4.3							
Instruction	N/A							
Argumentation	6.2.7							
Interaction	N/A							

Aptis Speaking

Categories Speaking	Task 1	(Task 1) Item 1	(Task 1) Item 2	(Task 1) Item 3
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus	giving personal information			
Task Level (CEFR)	A2			
Response format				
Planning time (y/n)				
Planning time (seconds)				
Pattern of interaction				
Nature of intelocutor input 1				
Nature of interlocutor input 2				
Informational functions (multiple selection possible)	8			
Interactional functions (multiple selection possible)	N/A			
Managing interaction functions (multiple selection possible)	N/A			
Content knowledge	1 (General)			
Cultural specificity	1 (Neutral)			
Features of the input	Features of the input	Features of the input	Features of the input	Features of the input
Description				
Domain	Personal	Personal	Personal	Personal
Nature of information	Only concrete	Only concrete	Only concrete	Only concrete
Topic	Daily life	Daily life	Daily life	Daily life
Presentation				
Features of the Response				
Expected ouput (description)		Spoken	Spoken	Spoken
Form		Spoken	Spoken	Spoken
Expected output (time in minutes)		0.5	0.5	0.5
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale	3.7.3			
Description	3.6.3			
Narration	N/A			
Exposition	N/A			
Instruction	N/A			
Argumentation	N/A			
Interaction	N/A			

Categories Speaking	Task 2	(Task 2) Item 1	(Task 2) Item 2	(Task 2) Item 3
Features of the TASK	**Features of the TASK**	**Features of the TASK**	**Features of the TASK**	**Features of the TASK**
Skill focus	Giving opinions/reasons			
Task Level (CEFR)	B1			
Response format				
Planning time (y/n)				
Planning time (seconds)				
Pattern of interaction				
Nature of intelocutor input 1	Scripted			
Nature of interlocutor input 2				
Informational functions (multiple selection possible)	8,2,3,6			
Interactional functions (multiple selection possible)	N/A			
Managing interaction functions (multiple selection possible)	N/A			
Content knowledge	1 (General)			
Cultural specificity	1 (Neutral)			
Features of the input	**Features of the input**	**Features of the input**	**Features of the input**	**Features of the input**
Description				
Domain	Personal	Personal	Personal	Personal
Nature of information	Fairly abstract	Mostly concrete	Fairly abstract	Fairly abstract
Topic	Culture and customs	Culture and customs	Culture and customs	Culture and customs
Presentation				
Features of the Response				
Expected ouput (description)		Spoken	Spoken	Spoken
Form		Spoken	Spoken	Spoken
Expected output (time in minutes)		**0.75**	**0.75**	**0.75**
CSE Descriptors	**CSE Descriptors**	**CSE Descriptors**	**CSE Descriptors**	**CSE Descriptors**
Overall scale	3.7.3, 3.7.1			
Description	4.6.2			
Narration	4.5.1			
Exposition	N/A			
Instruction	N/A			
Argumentation	3.2.3, 3.2.1			
Interaction	N/A			

Categories Speaking	Task 3	(Task 3) Item 1	(Task 3) Item 2	(Task 3) Item 3
Features of the TASK	**Features of the TASK**	**Features of the TASK**	**Features of the TASK**	**Features of the TASK**
Skill focus	Describe, contrast and compare			
Task Level (CEFR)	B1			
Response format				
Planning time (y/n)				
Planning time (seconds)				
Pattern of interaction				
Nature of intelocutor input 1	Scripted			
Nature of interlocutor input 2				
Informational functions (multiple selection possible)	2,3,4,5.6.8.10.11			
Interactional functions (multiple selection possible)	N/A			
Managing interaction functions (multiple selection possible)	N/A			
Content knowledge	1 (General)			
Cultural specificity	1 (Neutral)			
Features of the input	**Features of the input**	**Features of the input**	**Features of the input**	**Features of the input**
Description				
Domain	Personal	Personal	Personal	Personal
Nature of information	Mostly concrete	Only concrete	Fairly abstract	Mostly abstract
Topic	Descriptions of places (towns, cities, locations)	Descriptions of places (towns, cities, locations)	Descriptions of places (towns, cities, locations)	Descriptions of places (towns, cities, locations)
Presentation				
Features of the Response				
Expected ouput (description)		Spoken	Spoken	Spoken
Form		Spoken	Spoken	Spoken
Expected output (time in minutes)		**0.75**	**0.75**	**0.75**
CSE Descriptors	**CSE Descriptors**	**CSE Descriptors**	**CSE Descriptors**	**CSE Descriptors**
Overall scale	4.7.3, 4.7.1			
Description	4.6.2			
Narration	N/A			
Exposition	5.4.6			
Instruction	N/A			
Argumentation	5.2.7			
Interaction	N/A			

Categories Speaking	Task 4	(Task 4) Item 1	(Task 4) Item 2	(Task 4) Item 3
Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK	Features of the TASK
Skill focus	Long turn			
Task Level (CEFR)	B1			
Response format				
Planning time (y/n)				
Planning time (seconds)				
Pattern of interaction				
Nature of intelocutor input 1	Scripted			
Nature of interlocutor input 2				
Informational functions (multiple selection possible)	2,3,4,8,11			
Interactional functions (multiple selection possible)	N/A			
Managing interaction functions (multiple selection possible)	N/A			
Content knowledge	1 (General)			
Cultural specificity	1 (Neutral)			
Features of the input	Features of the input	Features of the input	Features of the input	Features of the input
Description				
Domain	Personal	Personal	Personal	Educational
Nature of information	Mostly concrete	Only concrete	Fairly abstract	Fairly abstract
Topic	Leisure and entertainment	Leisure and entertainment	Leisure and entertainment	Leisure and entertainment
Presentation				
Features of the Response				
Expected ouput (description)		Spoken	Spoken	Spoken
Form		Spoken	Spoken	Spoken
Expected output (time in minutes)		2	2	2
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale	4.7.1, 5.7.3, 6.7.2			
Description	5.6.3			
Narration	5.5.3			
Exposition	5.4.6			
Instruction	N/A			
Argumentation	6.2.3, 6.2.5			
Interaction	N/A			

Aptis Writing

Categories Writing	Task 1	(Task 1) Item 1	Task 2	(Task 2) Item 1
Features of the Task	Features of the Task	Features of the Task	Features of the Task	Features of the Task
Skill focus				
Task Level (CEFR)	A1		A2	
Response format				
Planning time (y/n)				
Planning time (seconds)				
Functions (BC EAQUALS list	5,6,9		18, 21,22,29	
Content knowledge	1 (General)		1 (General)	
Cultural specificity	1 (Neutral)		1 (Neutral)	
Features of the input	Features of the input	Features of the input	Features of the input	Features of the input
Description				
Domain	Personal		Personal	
Nature of information	Only concrete		Mostly concrete	
Topic	Daily life		History and archaeology	
Presentation				
Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response
Expected ouput (description)				
Expected output (number of words)				
Form of written output				
Discourse mode		Descriptive		Expository
Intended genre		Messages and short memos		Personal letters / e-mail
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale	N/A		3.7.2, 3.7.3	
Description	1.6.2		N/A	
Narration	1.5.2		3.5.1	
Exposition	1.4.2, 1.4.3		2.4.2	
Instruction	N/A		N/A	
Argumentation	N/A		N/A	
Interaction	N/A		N/A	

Categories Writing	Task 3	(Task 3) Item 1	(Task 3) Item 2	(Task 3) Item 3	Task 4	(Task 4) Item 1	(Task 4) Item 2
Features of the Task	Features of the Task	Features of the Task	Features of the Task	Features of the Task	Features of the Task	Features of the Task	Features of the Task
Skill focus							
Task Level (CEFR)	B1				B2		
Response format							
Planning time (y/n)							
Planning time (seconds)							
Functions (BC EAQUALS list	3, 5,18, 21,22,29				3,5.7.10, 11,18,21,22, 23,25, 29,31,38		
Content knowledge	1 (General)				1 (General)		
Cultural specificity	1 (Neutral)				1 (Neutral)		
Features of the input	Features of the input	Features of the input	Features of the input	Features of the input	Features of the input	Features of the input	Features of the input
Description							
Domain	Personal				Personal		
Nature of information	Mostly concrete				Mostly concrete		
Topic	History and archaeology				Leisure and entertainment		
Presentation							
Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response	Features of the Response
Expected ouput (description)							
Expected output (number of words)							
Form of written output							
Discourse mode		Descriptive	Descriptive	Expository		Narrative	Argumentative
Intended genre		Personal letters / e-mail	Personal letters / e-mail	Personal letters / e-mail		Personal letters / e-mail	Personal letters / e-mail
CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors	CSE Descriptors
Overall scale	3.7.3, 3.7.2, 4.7.1, 4.7.3				3.7.3, 4.7.3, 5.7.1, 6.7.3		
Description	N/A				N/A		
Narration	3.5.1				N/A		
Exposition	3.4.3, 4.4.1, 4.4.3				4.4.2, 4.4.3		
Instruction	N/A				N/A		
Argumentation	3.2.1				4.2.1, 4.2.2, 4.2.3. 5.2.2, 5.2.1		
Interaction	4.1.1				6.1.2		

Appendix D: Summary of questionnaire results

Listening

Gender	**Male**	**Female**
	7	9
First language	**Chinese**	**English**
	13	3

Professional experience	**Elementary school**	**Junior / Senior High School**	**University**	**Company / Business Classes**	**Other**
	2	6	15	7	8

Knowledge of the CSE and Standard Setting (tick the appropriate box)	**I had read the CSE and was familiar with its aims and content, including the Common Reference Levels.**	**I was familiar with the aims of the CSE, but had not studied it in detail**	**I had heard of the CSE but was not familiar with its aims or content.**	**I had not heard of the CSE.**
	8	5	2	0

Prior experience with standard setting	**I have had experience acting as a judge/rater on standard setting panels.**	**I have had experience organising standard setting panels.**	**I was familiar with the concept of standard setting,**	**I was familiar with the concept of standard setting, but had not heard of any of the methods used in the CSE project.**
	1	0	8	6

For each of the statements below, choose the option which most closely represents your opinion.				
	Strongly Disagree	**Disagree**	**Agree**	**Strongly Agree**
The preparation booklet gave me a clear understanding of the purpose of the project.	3	0	4	9
The explanations and tasks in the preparation booklet helped me understand the structure of the CSE.	3	0	5	8
The group discussion of the CSE at the start of the workshop aided my understanding of the CSE.	2	1	2	11
The explanation of the Aptis and IELTS tests was adequate.	2	1	4	9
The explanation of the Basket Method was adequate and I felt able to undertake the rating task.	3	0	3	10
The explanation of the Angoff Method was adequate and I felt able to undertake the rating task.	2	1	4	9
The time provided for rating the CSE level of items was adequate.	1	3	3	9
The feedback on all raters' judgements between rounds on participants' judgements was useful for making a final decision.	1	2	5	8
The feedback on actual item difficulty was useful.	1	2	9	4
During the seminar, I felt I had adequate opportunities to present my opinions and there was an equal opportunity for everyone to contribute his/her ideas during the discussion.	2	1	4	9
The facilities and food service were adequate.	2	2	4	8

Reading

Gender	**Male**	**Female**
	7	13
First language	**Chinese**	**English**
	16	4

Professional experience	**Elementary school**	**Junior / Senior High School**	**University**	**Company / Business Classes**	**Other**
	3	8	16	6	10

Knowledge of the CSE and Standard Setting (tick the appropriate box)	**I had read the CSE and was familiar with its aims and content, including the Common Reference Levels.**	**I was familiar with the aims of the CSE, but had not studied it in detail**	**I had heard of the CSE but was not familiar with its aims or content.**	**I had not heard of the CSE.**
	13	5	0	1

Prior experience with standard setting	**I have had experience acting as a judge/rater on standard setting panels.**	**I have had experience organising standard setting panels.**	**I was familiar with the concept of standard setting,**	**I was familiar with the concept of standard setting, but had not heard of any of the methods used in the CSE project.**	**I was not very familiar with the concept of standard setting**
	11	1	2	4	1

For each of the statements below, choose the option which most closely represents your opinion.				
	Strongly Disagree	**Disagree**	**Agree**	**Strongly Agree**
The preparation booklet gave me a clear understanding of the purpose of the project.	0	0	7	12
The explanations and tasks in the preparation booklet helped me understand the structure of the CSE.	0	0	9	10
The group discussion of the CSE at the start of the workshop aided my understanding of the CSE.	0	0	2	17
The explanation of the Aptis and IELTS tests was adequate.	0	0	4	15
The explanation of the Basket Method was adequate and I felt able to undertake the rating task.	0	0	2	17
The explanation of the Angoff Method was adequate and I felt able to undertake the rating task.	0	0	4	15
The time provided for rating the CSE level of items was adequate.	0	0	3	16
The feedback on all raters' judgements between rounds on participants' judgements was useful for making a final decision.	0	1	5	13
The feedback on actual item difficulty was useful.	0	0	5	14
During the seminar, I felt I had adequate opportunities to present my opinions and there was an equal opportunity for everyone to contribute his/her ideas during the discussion.	0	0	5	14
The facilities and food service were adequate.	0	0	6	13

Speaking/Writing

Gender	**Male**	**Female**
	6	**10**
First language	**Chinese**	**English**
	13	3

Professional experience	**Elementary school**	**Junior / Senior High School**	**University**	**Company / Business Classes**	**Other**
	1	6	12	2	12

Knowledge of the CSE and Standard Setting (tick the appropriate box)	**I had read the CSE and was familiar with its aims and content, including the Common Reference Levels.**	**I was familiar with the aims of the CSE, but had not studied it in detail**	**I had heard of the CSE but was not familiar with its aims or content.**	**I had not heard of the CSE.**
	15	4	0	0
Prior experience with standard setting	**I was a judge on the CSE standard setting panel for Aptis IELTS listening**	**I was a judge on the CSE standard setting panel for Aptis IELTS reading**	**I have been a judge on other standard setting panels**	**I had no experience of standard setting before this panel**
	10	10	5	3

For each of the statements below, choose the option which most closely represents your opinion.				
	Strongly Disagree	**Disagree**	**Agree**	**Strongly Agree**
The preparation booklet gave me a clear understanding of the purpose of the project.	0	0	3	13
The explanations and tasks in the preparation booklet helped me understand the structure of the CSE.	0	0	2	14
The group discussion of the CSE at the start of the workshop aided my understanding of the CSE.	0	0	3	13
The explanation of the Aptis and IELTS tests was adequate.	0	0	4	12
The explanation of the standard setting method was adequate and I felt able to undertake the rating task.	0	0	2	14
The time provided for rating the CSE level of items was adequate.	0	2	5	9
The feedback on all raters' judgements between rounds on participants' judgements was useful for making a final decision.	0	0	2	14
During the seminar, I felt I had adequate opportunities to present my opinions and there was an equal opportunity for everyone to contribute his/her ideas during the discussion.	0	0	3	13
The facilities and food service were adequate.	0	0	3	13

British Council Assessment Research Group

The Assessment Research Group was formed in 2013 to support the British Council's work in assessment and testing across the world. The team is responsible for ensuring that all new assessment products and new uses of existing products are supported by the most up-to-date research. They also continuously evaluate the quality of British Council assessment products.

TECHNICAL REPORT ON LINKING UK EXAMS TO CHINA'S STANDARDS OF ENGLISH LANGUAGE ABILITY (CSE)

VS/2019/003

BRITISH COUNCIL VALIDATION SERIES

Published by British Council
10 Spring Gardens
London SW1A 2BN

www.britishcouncil.org/aptis/research