BRITISH COUNCIL

ENGLISH LANGUAGE
ASSESSMENT RESEARCH GROUP

Technical Report

Speaking and Writing Rating Scales Revision

TR/2017/001

Judith Fairbairn, British Council
Jamie Dunlea, British Council

ISSN 2057-7168

www.britishcouncil.org/aptis

CONTENTS

LIST OF FIGURES

LIST OF TABLES

TABLES IN APPENDICES

1. INTRODUCTION

1.1 Purpose

This report describes a project to revise the rating scales used for the writing and speaking components of the Aptis test, which is a computer-based English proficiency test system developed by the British Council and launched in 2012. Aptis tests the four skills (reading, listening, speaking and writing) and was designed to provide cost-effective, efficient, and flexible testing options. For information on the Aptis test system, see O'Sullivan and Dunlea (2015), which provides detailed descriptions of the test design, task specifications and scoring.

1.2 Background

The Aptis test system was designed within the socio-cognitive theoretical framework model of test development and validation (O'Sullivan, 2011; O'Sullivan and Weir, 2011; Weir, 2005). This model focuses on collecting test validation evidence around three elements: the test-taker, the test system and the scoring system. This report is concerned primarily with the scoring system, specifically the rating scales used by human raters in the allocation of scores for the productive (writing and speaking) skill components of the test.

The productive skills tests each have four tasks and require test-takers to provide samples of spoken and written performances. The speaking test is a semi-direct test in which test-takers record responses to pre-recorded prompts. For the writing test, the tasks include writing emails, filling in forms and participating in online social media chat forums. The tasks and rating scales used to elicit performance are based on the socio-cognitive framework of language test development and validation.

Another important element of test design has been the explicit use of the Common European Framework of Reference (CEFR) (Council of Europe, 2001) guidelines into the Aptis test system. The CEFR is built into the scoring system, with test-takers receiving both a scale score and a CEFR level as feedback. The CEFR provides a descriptive proficiency scale covering six broad levels, which are further broken down into separate scales covering a number of areas of language use. Can-do statements, or descriptors, describe what learners at each level of proficiency are able to do with the language in each of these areas.

For the Aptis test system, the CEFR descriptors acted as a springboard for task design and rating scale development. The key words and concepts in the CEFR descriptors were modified based on the contextual and cognitive parameters in the socio-cognitive model. This approach has provided the means for creating detailed task specifications and rating scales (see O'Sullivan and Dunlea, 2015 for examples). The relationship between performance on each component of the test and reporting of CEFR levels was validated through a standard setting procedure (O'Sullivan, 2015b).

All writing and speaking tasks are marked by trained raters using rating scales within an online secure rating system. To apply for training and certification as an Aptis rater, prospective raters must have a language teaching qualification (such as a CELTA) as a minimum, experience working remotely and online, and experience using the CEFR. The raters train online for a recommended 20 hours over eight days and pass an accreditation test before starting live online rating from home. The online system replaced a face-to-face training model that was used during the initial stages of the Aptis system. The transition to an online training system took place over the same period as the scale revision project, and its development and validation is described separately in Knoch, Fairbairn and Huisman (2015; 2016).

1.3 Change within the Aptis test system

The theoretical framework for Aptis has, from the start, made flexibility and adaptation an explicit part of the test system. This openness to change is described by O'Sullivan (2015a, p. 4) in the following way: "The Aptis test system was designed to be dynamic and it is expected that revisions and changes to various aspects of the system will be implemented in the course of actively engaging with the needs of test users". An active research agenda facilitates the collection of validation evidence and the test system is responsive to this evidence when it provides indications that aspects of the test system could be improved. Aptis also has a system of localisation (O'Sullivan and Dunlea, 2015), which has facilitated the development of a number of variants of Aptis (e.g. Aptis for Teens, Aptis Advanced).

All tests are, to some extent, a balancing act between meeting the practical demands of the testing context and eliciting samples of performance through test tasks that are relevant to the construct to be tested. In the process of finding an optimal balance appropriate for the test purpose, compromises are always required. As we learn more about the particular context in which we are using a test, and as we evaluate the technical performance of our test, we can identify ways to refine and improve our test design. This project is a concrete example of this approach, and describes how evidence was collected on how the rating scales for the productive skills were being applied in practice with the resulting revisions to the scales.

2. RATIONALE FOR THE SCALE REVISION PROJECT

2.1 Initial scale development decisions

The first variant in the Aptis system, referred to as Aptis General, was launched in 2012. The productive skills components and scoring system were only part of the full test design leading up to the test roll-out. There were staff, including raters, to recruit and train; the online test system across all components to design; and a computer delivery platform to develop that could be used to test candidates across a range of global contexts with varying levels of internet and computer access. At all stages in this process, numerous requirements to balance the practicalities of a cost-effective online international test with robust testing theory were encountered.

In developing the scoring system for the productive skills, one of the most important decisions that needed to be made was in the approach to the rating scales to be used. One distinction commonly made is between analytic and holistic scales (for a detailed description of these two approaches, see the literature review in Section 3). For Aptis, it was decided to employ holistic scales rather than analytic scales. However, it should be noted that the holistic approach used by Aptis differs from the "impressionistic" holistic scales described by Weigle (2002), and which were the focus of many early comparisons between the advantages and disadvantages of the two approaches. The Aptis rating scales use a guided holistic scales approach, with analytic descriptions of performance at each level on the scale requiring raters to arbitrate between the criteria to decide on the CEFR level.

2.2 Different approaches for speaking and writing

The decision to use a holistic approach was an effort to balance practical aspects of rating with the need for reliable and accurate scoring that adequately reflected the performance of the test-takers. Holistic scales are quick and efficient – features which were emphasised in the design of Aptis to make it a flexible, cost-efficient assessment option. Another aspect that influenced the use of holistic scales was the level-specific nature of the four tasks. For the writing test in particular, there are four tasks, with each task targeting a performance relevant to a different level of the CEFR (Task 1 is targeted at A1, Task 2 at A2, Task 3 at B1, and Task 4 at B2). The nature of the performance targeted by each task has distinct features, resulting in clearly different rating requirements (e.g. register is a key criterion only for the B2 task). For this reason, it was decided to develop task-specific holistic scales for writing.

As described in O'Sullivan (2015a), it was further felt that there is a distinct difference in the processing demands between rating spoken and written performances. For writing, raters have access to the written performance, and can move easily back and forth over the response, revisiting and rechecking salient parts of the performance. With speaking, real-time processing is required for rating and the process of revisiting sections of the performance is much more difficult and time consuming. Because of this increased cognitive demand, and the increased time required when rating spoken performances, it was decided to use a single holistic scale for all four tasks in the speaking test. Although the speaking tasks also increase in difficulty, with each task targeting a specific CEFR level, the speaking tasks did not target clearly different aspects, such as register, for different tasks. It was felt the single holistic scale approach would adequately capture performance distinctions across tasks while maximising efficiency for the rating of that skill. The original scales were piloted prior to the launch of the test, and their performance was evaluated, not only from a technical scoring perspective, but also in terms of feedback on usability from raters. The link between the levels allocated by raters, the actual performance characteristics of test-takers' written and spoken responses, and the link to the intended levels of the CEFR was further validated in the standard setting study carried out before the test was launched (O'Sullivan, 2015b).

2.3 Monitoring operational use

With all test development it is essential to evaluate the technical performance characteristics of the test in operational circumstances. In the case of Aptis, which is delivered globally across a range of contexts by the British Council, it was also recognised that close monitoring would be required to note differences that might develop between the original piloting phases and actual operational use. For the writing and speaking tests, the growing global pool of raters offered the opportunity to collect evidence of how raters interacted with the scales and to provide insight into the adequacy of the fit between the rating scale descriptors and test-taker performance. Such a process of review and potential revision was in line with the design concept of the Aptis test system noted in Section 1.3. The system would be dynamic, responsive and open to change when sufficient evidence was available.

After the initial launch of the test, a number of opportunities were provided for monitoring the operational usability of the scales from the raters' perspective. At the same time, periodic recruitment and training sessions to increase the rater pool also provided the opportunity to discuss the use of the scales in relation to the standardised exemplars of performance, selected from live tests, used in training. Test administration colleagues were also in contact with the examiner network manager to resolve any questions raised by test-takers regarding results. While much of this information was individual and anecdotal, some evidence of certain trends and issues in the application of the rating scales began to appear.

2.4 Issues raised by raters

Raters would make the same comments about the rating scales, even after additional training, indicating where the rating scales might be unclear. For example:

- Task completion was not included in the rating scales, and this led to some variation in the interpretation of how well test-takers had fulfilled specific task requirements. This was particularly evident with tasks that contained a set number of questions. Test-takers sometimes varied in the number of questions they addressed directly, with some test-takers addressing fewer questions, but also providing extended answers for those questions.
- Raters appeared to have some difficulty using the single, holistic scale which had been developed for use with all four speaking tasks. The limitations in the required output for the A2 task demands meant that test-takers who might be able to demonstrate B2/C1 performance features on the B1 and B2 tasks would not demonstrate them on the A2 task.
- Raters noted some difficulty with marking exceedingly short or long writing texts.
- Poor recording quality was at times mistaken for poor pronunciation.
- Some measures of fluency, such as intonation, were being marked inconsistently.
- There was a reluctance to give top scores to test-takers who were clearly above B2 level.

Most of the evidence collected was qualitative, subjective evidence, often relying on the expert judgement of raters and the examiner network manager or senior raters mediating questions raised by test-takers or test centres over individual scores. Therefore, after one year of using the original rating scales and collating indications of these trends and issues with the scales, the Assessment Research Group (ARG) decided to review the scales more formally.

3. LITERATURE REVIEW

3.1 Introduction

A rating scale is normally used to rate the productive speaking and writing skills and represents the test construct, or the components of the written or spoken performance, to be measured. The scale usually has descriptors of the performance that each test-taker level is expected to achieve, and raters place the test-taker at a level on the rating scale. Marking the full writing or speaking construct is subjective and more complicated than making 'right-wrong' decisions (Alderson, Clapham & Wall, 1995; McNamara, 1996).

3.2 Rater effects

The subjectivity of the marking creates rater effects (Myford & Wolfe, 2003). Researchers categorise rater effects differently and disagree about how to analyse the data but the main rater effects are:

- **leniency/severity** where raters rate too high or too low
- **inconsistency** where raters apply the rating scale in a different way to what is intended
- **halo effect** where raters are unable to distinguish between different categories and allocate similar scores to everyone
- **central tendency or restricted range** where raters avoids extreme ratings or one part of the rating scale
- **bias** where raters mark a particular group of people in a particular way
- **logical errors** where raters mark related features of the speaking or writing performance in the same way
- **basic errors** where raters make marking mistakes perhaps due to fatigue.

These rater effects cause construct irrelevant variance, in that the features of the spoken or written performance (e.g. grammar, lexis, cohesion) are not being measured as intended by the test developer because the rater is not following the construct expressed in the rating scale and, therefore, test construct validity is reduced (Huot, 1990b; Myford and Wolfe, 2003). The impact of rater effects on score validity is such that inferences made about a test-taker's ability may, therefore, be false (Bejar, 2012). If the rater effects are widespread, perhaps because the test is difficult to mark, test validity decreases and we are unable to make arguments about the proposed interpretations or uses of the test scores (Kane, 2013). It is, therefore, important to have quality assurance practices in place to ensure acceptable rater reliability so that that test results can be generalised; success on a test must demonstrate success in the tested skill (Huot, 1990b).

One way of ameliorating the impact of rater effects is the use of multiple marking. Multiple marking, however, has cost implications so to maximise cost-efficiency for test users and to turn around results quickly, the decision was taken to adopt a single-marking approach. As already noted, all test systems have to make choices to reach an optimal balance of the features relevant to a specific context of use. For Aptis, the single-rating decision, therefore, required consideration of alternative methods of quality assurance to help maintain the balance between practical considerations for efficiency and acceptable quality assurance in terms of rating consistency and accuracy. Several technological features of the computer delivery system were exploited to help achieve quality assurance. Firstly, the system allowed the test-taker performances to be broken up and distributed to different raters, bringing the different task ratings back together for the final results calculation. Using this feature, the four task performances for the same candidate on both the speaking and writing components could each be marked by a different rater, therefore reducing the impact of rater effects on any one test-taker. Secondly, Aptis was able to implement a system of control items. Interspersed within the live items marked online are control items or "gold standard seeding" (Shaw and Weir, 2007). Control items (CIs) are benchmark candidate performances that have been marked by a group of experienced markers, which are given to raters at the start of their marking session, and then randomly at a rate of 5% (1 in 20). Raters are aware that they will be presented with CIs, but there is no distinction in presentation between CIs and operational live marking responses. This system automatically suspends raters who are marking outside of a set tolerance for a given task and can provide operational estimates of rater reliability (O'Sullivan and Dunlea, 2015).

3.3 Rating scales

On top of good quality assurance processes, well-designed rating scales are essential and can help mitigate rater effects. A rating scale with a variety of interpretations will cause raters to interpret the criteria differently and reduce rater reliability. The rating scales can be analytical in that each feature of speech or writing is marked independently and can be weighted, or the scales can be holistic and the test-taker is placed in a category where their performance as a whole best fits, based on blended criteria (Hamp-Lyons, 1995; Huot, 1990a, 1990b; Weigle, 2002). There is also primary or multiple trait scoring which involves the rater identifying one or more specific features of the response.

There are benefits and drawbacks to different rating scales and the choice depends in part on the assessment purpose (Weigle, 2002). Analytic rating can more easily detect if a particular linguistic feature is generating a rater effect because each feature is marked individually, and capturing feedback on each feature is diagnostically quite useful for test-takers (Hamp-Lyons, 1995). Holistic rating scales combine descriptions of features and thus make the score difficult to interpret, but might be adequate for a placement test, and is much quicker to mark (Weigle).

Marking separate features analytically focuses the attention of the rater on the marking construct and, therefore, may lessen the cognitive load of having to weight different features, as a rater needs to do when marking holistically (Weigle, 2002). Inexperienced raters in particular might find it difficult to internalise the differences between holistic score levels and find it easier to focus on individual criteria. Barkaoui (2011), who trained raters to use analytic and holistic writing rating scales and then had them mark the same tests using each system, found more inconsistent marking for inexperienced raters when marking holistically. The inexperienced raters did not use the full rating scale and tended to allocate similar marks around the central value to all test-takers (central tendency rater effect), suggesting that holistic rating may require a higher level of expertise and practice to master the cognitive load requirement to arbitrate between levels. While Weigle and Barkaoui find analytic rating to be cognitively less demanding, Bejar (2012) hypothesised that, because there are more decisions to make and more scores to decide, the cognitive load may actually be higher for analytic rating. Seedhouse, Harris, Naeb and Ustunel (2014) also note the high cognitive load in trying to match the response to each analytic rating scale. Further research is needed into which rating scale has a higher cognitive load but the answer could be related to the decision-making cognitive style preference of individual raters.

Research is mostly focused on the writing skill and the findings are mixed as to whether analytic or holistic rating is more reliable. Some researchers find that, with more decisions made in analytic rating, there will be more diversity of opinions and therefore less inter-rater agreement (Bejar, 2012). Barkaoui (2011) also found that, although analytic rating had a more detailed analysis of the test-taker performance due to the separate scores for each feature, the increased number of decisions led to lower inter-rater reliability for the final mark. Interestingly, Barkaoui found that holistic marking led to lower intra-rater reliability; raters were less internally consistent with themselves when marking holistically. Perhaps without a clear focus, which raters get when marking analytically, the rater might change their focus from one marking session to the next. By contrast, Huot (1990b) found analytic rating to have higher inter-rater reliability and so did Cumming (1990). The research findings are, therefore, mixed on which scale has higher reliability, but it could be in part dependent on the quality of the rating scales.

One constraint in Barkaoui's (2011) research was possibly the rating scales. The same marking criteria and wording were used for the analytic and holistic rating scales. Raters marked the same writing scripts two weeks apart using each rating scale. For those who first marked analytically, there may have been a temptation to repeat the analytic decision-making process when marking holistically. Or the opposite could have occurred. Cumming (1990) found that raters tend not to vary their marks across the different analytic categories, focusing instead on a final holistic mark and adjusting the analytic marks accordingly. The raters who started with the holistic marking in Barkaoui's research may have had a fixed final level in mind when marking analytically. Barkaoui's raters ranked the test-takers in the same order using both rating scales, indicating that the same construct was measured using analytic and holistic scales. Sweedler-Brown (1985) also had this finding but only with experienced raters.

One issue with holistic rating scales is that, even though they can produce reliable results, the exact constructs being assessed are unclear (Weigle 2002). Raters might agree in terms of an overall holistic mark but disagree on the reasons for the mark. Papajohn (2002), working on the speaking skill, looked at how raters derive scores and found a wide range of approaches. Raters were asked to design a 'concept map' of the decisions they make when marking, which they could approach in whichever way they thought appropriate. Some raters had a series of steps, others used Likert scales for different features and others used yes/no paths. Some focused first holistically and then funnelled into the detail. Others started with the detail and funnelled out. Therefore, it cannot be assumed that raters approach rating scales in a similar manner. These diverse decision-making preferences may impact on rater reliability.

Huot (1990b) raises the point that holistic scoring may be more valid than analytic scoring because it is a more authentic method of assessing communication. The sum of the analytic parts does not necessarily equal the whole response because it isolates linguistic features from context (Goulden, 1994, as cited by Barkaoui, 2011). However, this authenticity can influence the rating focus. The holistic rater may not notice individual salient features of a performance and instead focus on task fulfilment, content, or organisation of the response, as one would do in real life. Moreover, the focus could be completely different for each rater depending on their personal cognitive style.

3.4 Rater cognition and decision-making

Research shows that, even with robust rating scales and careful training, there is still marking inconsistency. An area of growing interest in language testing is rater cognition and the decision-making process. The understanding and acceptance of test constructs and rating scales requires a certain level of cognitive ability and some rater variability may be due to individual cognitive differences (Baker, 2012). The rater must be able to link the scoring criteria to the response being marked and the quality of the rating depends on how well the rater can make this link. For the test to have score validity, the rater's cognitive style for decision-making must also be consistent with the construct of the test (Bejar, 2012).

The rater's cognitive style is linked to their personality (Messick, 1994) and fits into a decision-making strategy type developed by Scott and Bruce (1995) as outlined by Baker (2012) and Spicer and Sadler-Smith (2005):

- Rational: structured collection and evaluation of information
- Intuitive: reliance on feelings, hunches, and impressions that cannot be put into words
- Dependent: receiving second opinions, direction, advice or support from others
- Avoidant: postponing, hesitating or avoiding decision-making
- Spontaneous: coming to an impulsive decision immediately or as early as possible.

Raters may straddle a few different decision-making strategies or have one strategy that dominates (Spicer & Sadler-Smith, 2005). Personality characteristics, such as self-confidence and levels of anxiety and interest, also feed into the rater's cognitive style (Myford and Wolfe, 2003). A rater's cognitive style for decision-making may impact on a preference for analytic or holistic rating scales. As mentioned earlier, it is unclear if the cognitive load is higher for analytic or holistic rating and it could be that the rater's cognitive load using each marking method is related to their cognitive decision-making preference.

Zelniker and Jeffrey (1976, 1979), as cited by Messick (1994) found that reflective and strategic individuals tend to analyse information in component features, whereas impulsive and non-strategic individuals are better at treating the information as a whole. It would be reasonable to conclude that rational, strategic and reflective raters would work better with analytic scales and intuitive, non-strategic, impulsive and spontaneous raters would prefer holistic scales. In practice, however, statistical significance between cognitive style decision-making preferences and rating accuracy has not been proven, probably due to the limitations of self-reporting by raters who tend to overrate themselves as rational (Baker, 2012). More research is needed in this area perhaps using eye-tracking software to determine how raters approach rating scales, an approach which has been used successfully to provide insights into test-taker cognitive processes in a number of studies (for example, Bax, 2013 and Brunfaut et al., 2015 in relation to reading; Holsknecht et al, 2017 with listening).

Thunholm (2004) queries whether decision-making strategies are stable or easily changed. As mentioned already, the decision-making style must be consistent with the test construct, so it is important that raters are able to work within the decision-making style of the test. For some raters, a switch away from their preferred decision-making style may cause additional cognitive load. This is another interesting area for further research, including how raters change from different rating scales for different tests.

Other rating strategies include placing high or low importance on a particular linguistic feature (Eckes, 2012). Brown, Iwashita and McNamara (2005) found that linguistic resources (grammatical and lexical range and accuracy) are focused on the most when marking speaking. Cumming, Kantor and Powers (2002) found that raters mark on ideas instead of language. Raters also use criteria not mentioned in the rating scales and also differ in their views of what constitutes fulfilment of the criteria (Brown et al., 2005; Weigle, 2002). Brown et al. and Cumming et al. found that raters focus on different areas depending on the task type. The choice of task type, task topic, and task difficulty may favour raters with specific cognitive decision-making styles (Baker, 2012; Eckes, 2012; Huot, 1990b).

Rater experience may also interact with a rater's cognitive style and impact on reliability (Baker, 2012; Bejar, 2012) but the research is mixed. Experienced raters appear to be able to reflect more on the differences between the rating levels and highlight linguistic accuracy, while inexperienced raters tend to highlight linguistic errors and have more misfit, i.e. they are less accurate (Barkaoui, 2011). Sweedler-Brown (1985) also found inexperienced raters to be less accurate while Lim (2011) and Fairbairn (2015) both found that inexperienced raters were less accurate but that their accuracy improved quickly. Weigle (1998) also found inexperienced raters to be less accurate before training, and more severe, but that the differences between the two groups of raters were less pronounced after training. Cumming (1990), by contrast, found inexperienced raters to be more lenient. Huot (1990b) found no difference between new and experienced raters, but noted that experienced raters appeared to have more efficient rating processes.

In Alderson's (1993) review of the literature on judgements made in language testing, he noted that professional judgements by testers are frequently in conflict. He was looking at test content, test and item difficulty and decisions on grade boundaries, and not specifically at rating, but the findings are still relevant to this discussion. In the study on test content, for example, he found that testing professionals disagreed about which feature and level of language were being tested, which shows that raters may come to a testing situation with their own views on the test construct, which may not be the same as the test developer. As Papajohn (2002) noted, raters have their own internal rating culture and habits. The importance of solid rater training cannot be highlighted enough.

4. RESEARCH METHODOLOGY

4.1 Defining the scope of the project

At the outset of the project, the scope of the revision was defined as focusing on improving the clarity and usability of the rating scales and to place task design outside of the scope of this project. The test had been piloted in the development stage, and operational testing had been underway for a little over a year. The feedback from test users had been generally positive regarding the actual format and design of the test tasks, and the issues that had been raised related to the scoring system rather than the test system. At the same time, revision of the test tasks in additon to the rating scales would have entailed not only a much larger project, involving the design, development and trialling of new task types, it would have required a much wider coodination of resources across marketing and general communications to ensure that all test users were aware of the changes. For these reasons, it was decided that a restricted focus on improving the scoring validity of the rating scales was both practical and justified for this stage in the operational life-cycle of the test.

4.2 Developing a cyclical process of scale revision

The scale revision project took one year and followed a series of steps illustrated in Figure 1 (adapted from Dunlea, Fairbairn, O'Sullivan, 2016). The data collection instruments included the following:

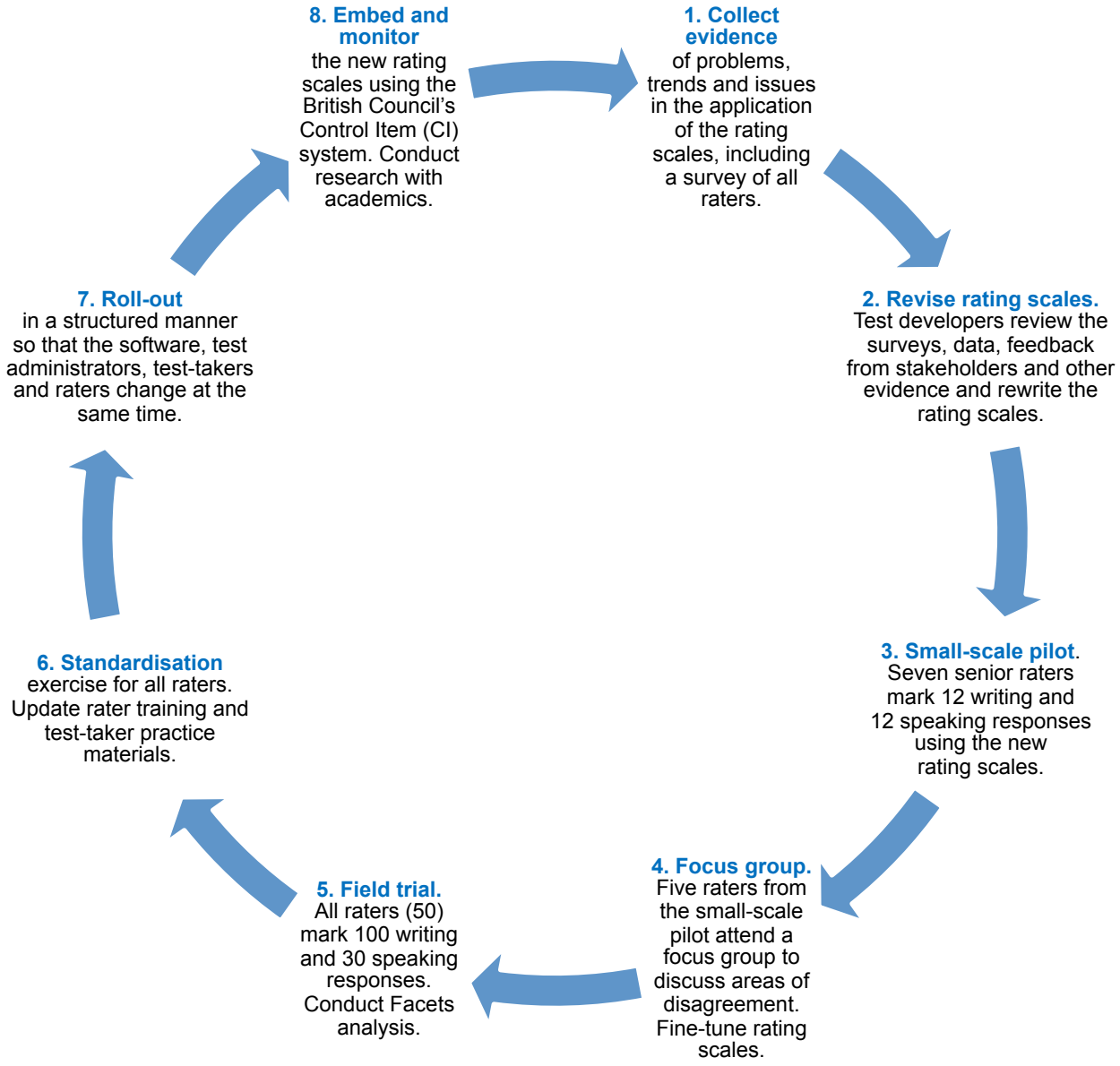
- questionnaire to all raters
- focus group with assessment experts
- small-scale pilot of new rating scales
- focus group with small group of raters
- field trial of rating scales with all raters.

The cycle reflected in Figure 1 is a part of the ongoing attempt to develop systematic approaches to language testing research and validation that are operationally relevant and grounded in robust theory. On a much smaller scale, Figure 1 reflects the retrospective distillation of the varied activities of the TOEFL iBT development into the validity argument framework discussed by Chapelle et al. (2008). As Chapelle et al. (2008, p. 23) state:

> "The linear structure of the TOEFL intepretative argument that we have formulated in retrospect fails to capture the dynamic process that went into its construction…given the various types of evidence that can be offered in support of test interpretation and use, this interpretative framework now provides us with a way to organize the evidence and its implications. An argument-based approach was used to help collate and interpret the evidence already collected, not to drive the collection of that evidence."

Figure 1 was not developed *a priori*, but was produced as part of the reflection and evaluation of the project carried out after the revised scales had been launched and brought into operational use. In doing so, the authors of this report, in conjunction with the Assessment Research Group, were hoping to develop clear and explicit models which would ensure the work of collating, evaluating and acting on evidence of areas for improvement in the test system, which would become an iterative and established part of operational practice. The steps in the process are each explained briefly below.

Figure 1: Cyclical Aptis scale revision model



4.2.1 Collect evidence to inform revisions

As noted above, the collection of anecdotal evidence regarding the usability of the scales was an ongoing part of the examiner network manager's activities. As the evidence of certain trends began to emerge, the collection of this evidence took on a more systematic form, and anecdotal feedback of problems, trends and issues in the application of the rating scales was collated. A few test development and localisation research projects had been carried out during the same time period, which involved multiple rating and this data also added statistical insight into which tasks and types of responses seemed to be confusing to mark.

This process helped to identify potential areas of concern, and these were put into a questionnaire and administered to a wider sample of raters (see Appendices 2 and 3). Thirty-eight (38) out of 50 raters responded to the questionnaire. Some of the anecdotal evidence was substantiated with the questionnaire responses and other interesting findings were uncovered.

4.2.2 Revise rating scales

Next, a project group was formed consisting of members of the Assessment Research Group (ARG) and the Aptis test production team in charge of commissioning item writers for the production of writing and speaking tasks for operational tests. This project group met to review the questionnaire data and feedback from other stakeholders such as testing centres, test-takers and clients. The purpose was to make initial recommendations for changes to the rating scales. The project group agreed on the scope and focus of the revision project and tasked a core working group (consisting of the authors of this report) with the process of following up on the initial trends identified from the collated anecdotal, questionnaire, and test development evidence.

The core working group looked at the existing scales and identified areas for revision. This process started with a thorough review of the CEFR can-do scales for speaking and writing relevant to the Aptis test. As Aptis is an online test with a semi-direct format in which test-takers respond to pre-recorded prompts, some aspects of the speaking and writing construct were not being tested, such as interaction. Descriptors were chosen that matched the construct being tested in each task. The rating scale development process was iterative over several months and resulted in task specific rating scales focusing on the target CEFR levels and the rating would be a holistic two-stage range-finding process.

4.2.2.1 Task specific rating scales

At the review of anecdotal and questionnaire evidence carried out by the project group, it was agreed that the available evidence indicated that the task specific holistic scale approach for the writing test was working appropriately. On the other hand, the single holistic scale for the speaking test across all four tasks had shown some issues, including under-marking for the A2 task (see Section 2 for more information). The project group agreed that the initial rationale for a distinction between the scales for writing and speaking did not seem to be supported by the evidence from operational use. Therefore, it was agreed to change the approach for speaking, shifting from the original design decision to have one rating scale for all of the speaking tasks to an approach similar to the writing test with separate task-specific scales. The possible risk that the cognitive load would be too high with so many different rating scales was not supported by the questionnaire data and statistical information from multiple rater studies.

4.2.2.2 Developing a two-stage range-finding process for raters

There are four speaking and four writing tasks in the Aptis test. Table 1 outlines the target levels for each task. Writing Task 1 (A1) does not have a rating scale because these are short answer responses; Speaking Tasks 2 and 3 both target B1.

Table 1: Aptis Speaking and Writing target CEFR levels for tasks

Speaking		Writing	
Task	**Level**	**Task**	**Level**
T1	A2	T1	A1
T2	B1	T2	A2
T3	B1	T3	B1
T4	B2	T4	B2

It was decided that Tasks 1–3 for both skills should have 6-point rating scales (0–5) and Tasks 4 should have a 7-point rating scale (0–6). The target area would be a score of 3 or 4, which means that, at these scores, the test-taker has demonstrated sufficient performance at the CEFR level for the task (i.e. a 3 or 4 on Writing Task 3 means that the test-taker has produced B1 level writing). A 5 indicates a performance likely to be above the target CEFR level and a score of 0, 1 or 2 is below. For Task 4, (B2 tasks) a mark of 5 indicates a mark above B2 level. A mark of 6 was also included for a C2 response. A C2 level is not reported in the final score report in the Aptis General test – the 5 and 6 marks are conflated to a mark of C for the test-taker. Speaking Tasks 2 and 3 both target B1 level and would have the same rating scale. See Appendix 1 for an example of an Aptis rating scale. The full set of marking scales can be found in O'Sullivan and Dunlea (2015).

To help streamline the rating process and reduce the cognitive load on raters, a two-part range-finding holistic rating process was developed. The scales would take the advantages of the quick holistic marking system but add in a second decision in order to mitigate some of the problems associated with holistic marking. The rater would first mark the test-taker's response (the written or spoken production) holistically based on the qualitative linguistic features of the test-taker's performance (e.g. grammar, vocabulary, cohesion). The rater would place the test-taker at a CEFR level based on the overall language sample using the descriptors in the rating scale. The test-taker might have some features which are stronger or weaker than others and the rater would need to arbitrate between levels and linguistic features and decide which level the test-taker should be placed overall as a best fit.

The rater then would decide if the test-taker has sustained the CEFR level across the response. For example, Writing Task 3 targets B1 level. This means that a test-taker at B1 level should be able to adequately respond to the task. If the test-taker is placed at B1 level based on the qualitative linguistic features of their performance, the rater then decides if they have been able to sustain the B1 level throughout the response. They are marked at a lower B1 level if they are not able to sustain the B1 performance (i.e. B1.1 or a score of 3) and a higher B1 level if they have been able to sustain the B1 performance (i.e. B1.2 or a score of 4). If they meet all the criteria for B1 level, and exceed B1 level in at least one area, they are awarded a 5 (above B1 level). The test-taker may be placed at A2 level and the same decisions about their ability to sustain the response at A2 level are made for marks of 1 or 2. A mark of zero means that the test-taker is below A2 level.

This marking method was chosen because the distinction between CEFR levels for linguistic features was often very difficult to determine in an operational context. Therefore, it was decided to keep the qualitative descriptions of performance the same for two score points and then focus on the ability to sustain the level throughout the response as a way to distinguish between high and low level test-takers (i.e. B1.1 and B1.2 have the same qualitative descriptions of performance and differ only on the ability to sustain the performance).

4.2.3. Small-scale trialling

Once the project group was in agreement with the new rating scales, a small-scale trial of the new scales was conducted. Seven senior raters marked 12 writing and 12 speaking responses using the new rating scales. Although the pilot began with eight raters, one rater demonstrated a marked difference in rating patterns, and in discussion, indicated a clear misunderstanding of the activity, so this rater was dropped from the data used for this stage of the project. The raters also filled in a questionnaire to provide feedback (see Appendices 2 and 3). Further revisions were made to the wording of the scales to reflect what was learned from the raters' questionnaire feedback and marking performance in this small-scale trial.

4.2.4 Focus group

From the small-scale trial, five of these raters attended a face-to-face focus group to discuss areas of agreement and disagreement (two raters were not able to attend). The meeting was recorded and minutes were taken of key points. A document of frequently asked questions and definitions was also developed to be used later for training the entire rating cohort. The updated scales were discussed and checked to ensure that all issues raised during the focus group had been addressed. The rating scales were further fine-tuned after the focus group.

4.2.5 Field trial

Following the small-scale trialling of the revised scales, focus group and subsequent further revisions, a large-scale field trial was carried out with all raters operationally active in the international pool at the time (49 people) using Moodle. Raters marked 100 writing and 30 speaking tests after a short training course. The data was analysed using multi-faceted Rasch analysis. The results showed that the rating scales were generally working as intended. Some small refinements to the wording of the descriptors in the rating scales were made following the field trial.

4.2.6 Standardisation

Raters next underwent a standardisation exercise prior to the roll-out of the new scales. Test-takers were also informed about the new rating scales in the practice materials on information for candidates. The training course for new raters and test-taker practice materials were also revised.

4.2.7 Roll-out

The roll-out was conducted in a coordinated manner with raters, test-takers, test administrators and the scoring algorithms changing at the same time.

4.2.8 Imbed and monitor

The first six months after the new scales were in place were focused on embedding the new scales and making sure raters were marking to standard.

5. RESULTS

5.1 Overview

This section describes key results from the initial feedback questionnaire to examiners, the small-scale pilot study including the questionnaire and rating correlations, and finally the large-scale field trial, focusing on the multifaceted Rasch analysis.

5.2 Rater questionnaire

Data collected in the initial questionnaire to all raters (38 responses) helped identify which areas of the rating scales were problematic (see Appendices 2 and 3).

5.2.1 Writing

The A1 short-answer response task was not included in the study as it is a very limited form-filling task in which test-takers write only simple words or phrases to complete a form. At the time, the marking for this task did not use a rating scale with descriptors covering different aspects of performance, but instead focused on the number of gaps correctly filled.

The second task in the Writing component, which targets A2-level performance, is a 20–30 word short constructed response to a specific question. Test-takers are expected to respond with several short sentences to demonstrate A2-level performance. Raters were satisfied with the A2 writing rating scale with most agreeing or strongly agreeing that the rating scale was useful and clear with an appropriate number of marking points. Only eight of the raters would prefer to mark the A2 task analytically (see Appendices 2 and 3).

The most important criteria used to mark the A2 task was ‘overall impression’ and ‘relevance of content to the topic’; the least used criteria was ‘register’.

For Task 3, a B1-level writing task, the result was similar. The B1 task is a chat room with three questions and the expected output is 30–40 words per response. Raters were even more satisfied with the B1 task rating scale and again did not express a preference to mark analytically. For this task, ‘Relevance of content to the topic’, ‘task completion’, ‘grammatical accuracy’, and ‘vocabulary accuracy’ were chosen as being the most important. Register was again the lowest used criteria.

Task 4, the B2-level task, consists of two emails, one to a friend and one to a person of authority. The B2 task explicitly targets the ability to demonstrate control of register in the two emails, and again showed similar findings. Raters were satisfied with the rating scale and did not want to mark analytically, although more examiners did express a preference for marking analytically than in Tasks 2 and 3. For this task, however, all 38 raters agreed or strongly agreed that all performance features were important.

5.2.2 Speaking

As noted above, Speaking originally had a single holistic rating scale for all four tasks. Examiners strongly or very strongly agreed that the rating scale was useful for evaluating performances on all four tasks and again did not prefer to mark analytically. ‘Overall impression’ was the highest criteria used to mark and ‘intonation’ was the lowest. Raters noted in the comments section that the difference between the marks on the speaking scale were difficult to distinguish, particularly in the mid-range.

5.2.3 Rating behaviour

Although the rating scales are holistic scales, each band contains distinct aspects that could also be presented separately as analytic scales. One area of interest in rating behaviour was to investigate whether raters actually addressed the descriptors in the band holistically by evaluating which band provided the best fit for the performance, or whether raters in fact considered each separate aspect individually as if they were in fact analytic scales. Raters mostly always or usually read the task and task instructions before marking and did not rely on their first overall impression when deciding the rating. The questionnaire seems to demonstrate that raters did not, in fact, treat the separate strands of each band descriptor analytically and then average the scores to derive a final task rating.
See Appendix 3 for the questionnaire responses.

5.3 Small-scale trial

Raters marked 12 responses for each task type in each skill, filled in a questionnaire and participated in a face-to-face focus group meeting. The questionnaire to the seven raters involved in the small-scale trial of the new rating scales can be seen in Appendix 4 and the questionnaire responses are in Appendix 5.

The new scales were generally well-received, although the difficulty in changing to new scales, specifically for speaking which changed from one scale to three, was noticeable. The main difference in the rating scales was that now, instead of one rating scale for the four speaking tasks, there would be three different rating scales targeting the A2 task, the two B1 tasks and the B2 task, and this change took some getting used to. The focus group was invaluable in clarifying the reasons for divergences in marking, identifying confusing descriptors and collecting a set of FAQs for later dissemination to all raters.

5.3.1 Rating correlations

The ratings were analysed for marking similarities and differences and inter-rater correlations for each task were calculated. Average correlations were high for the writing scales, as would be expected as the changes had been minimal, but overall they were quite low for speaking, except for Task 3 (see Table 2). The responses with divergent ratings were discussed in detail in the focus group. Part of the goal of the focus groups was to dig into reasons for unexpected interpretations and divergent responses in order to, not only refine the wording of the scales, but also to develop guidelines and training procedures to help raters understand the intended use of the scales. This was particularly important with speaking, where the changes had been more comprehensive.

Table 2: Inter rater correlations for small-scale trial

Task	Average correlations
Writing Task 2	0.856
Writing Task 3	0.844
Writing Task 4	0.824
Speaking Task 1	0.499
Speaking Task 2	0.322
Speaking Task 3	0.757
Speaking Task 4	0.331

5.2.3 Questionnaire

The questionnaire feedback from the focus group members indicated that they generally had a favourable impression of the new rating scales, found them more detailed, and that they gave a more 'accurate picture of a candidate's performance'. Raters had more difficulty with the speaking rating scales because these scales had undergone a much larger change and there was some reluctance to change; in particular, one rater who found it very difficult to use the different task-specific scales for speaking, which partially explains the low average correlations. Raters also commented on the improved bullet-point layout, which made the salient features within each score band more transparent and readily accessible.

The questionnaire also asked raters how a writing text that was too long would impact on their marking. There were various strategies mentioned such as focusing only on the first 30 words (for the A2 task), seeing a long script as a sign of fluency and marking higher or alternatively exposing more errors and marking lower; however, the majority said that long scripts would have no impact. These responses identified the importance of providing examiners with instructions on how to deal with long responses that exceed the word count in the instructions to test-takers.

Another aspect investigated in the questionnaire was how raters would mark partially inaudible tests for speaking. Raters noted that because the new rating scales focused more on task achievement and the ability to sustain the CEFR level throughout the full response for each task, there could be more responses that examiners would escalate as inaudible responses. These responses would need to be reviewed by a senior examiner who could look at all four tasks to determine the final score, or ask for a retest if necessary.

One addition to the rating scales was to extend the scale for the B2 tasks (Tasks 4) to a 0–6 scale. The tasks target B2 level, with the 3 and 4 bands describing the target B2-level performance sufficient for the task. Bands 5 and 6 contain descriptions of performance relevant to C1 and C2 level. The raters expressed a positive attitude to this addition. Questions were raised as to whether a task targeting B2 could elicit C1 and C2 language. From the project group's perspective, it was felt that as the B2 task is an extended production task, test-takers would have the potential to demonstrate higher ability. At the same time, one of the issues associated with the scale usage after initial implementation had been reluctance by raters to award high scores, associated with a perceived need for high scores to be a 'perfect' response. In fact, each task was targeted at a specific CEFR level, and performance sufficient for that level should be enough to achieve the 'target' bands of 3 or 4 for each task-specific scale. A 5-point response on each scale would allow raters to identify responses that are distinctly beyond the target level, but nonetheless may not be 'perfect'. For B2 tasks, adding a 6-point band created the conceptual space for raters to be able to identify responses distinctly above C1, and to further create a psychological safety valve which would allow raters to award a 5 for C1-level performances, while still recognising that some improvement and sophistication in response could occur beyond that (which would be captured by the 6-point band).

In addition, feedback from a number of projects had requested the ability to distinguish between performances at these levels, and so it was recognised that in project-specific contexts, the B2 task was eliciting a range of performances at a high level, and the new scales would allow for distinctions to be made where necessary.

In practice, this ability to distinguish between high-level performances remained 'behind the scenes' as far as test-takers were concerned, as task-level scores are not reported for Aptis writing or speaking components. When determining and reporting the CEFR level for Aptis General components, the top level reported would remain as 'C', designating a strong performance likely to be above B2 (see O'Sullivan and Dunlea, 2015 for a full description of the scoring and reporting for Aptis tests). The scales, however, provided the ability for project-specific situations for performance distinctions to be made at a high level, and the 6-point scale was adapted for use in the development of an Aptis Advanced variant with higher level tasks which does report at C1 and C2 level.

5.3 Field trial

5.3.1 Overview

The field trial involved 49 raters operationally active at the time. A total of 100 writing test performances and 30 speaking test performances, representing a range of CEFR levels, were obtained from live tests delivered in operational situations. All tests had been marked under operational conditions, and the original scores awarded using the original scales were used as an initial indication of the level of the performance. Performances were selected based on the CEFR level allocated at the test level. A roughly equal number of performances were selected from test-takers who had achieved C, B2, B1, A2 and A1 level classification on the writing and speaking tests. Performances were not selected that had been allocated A0 overall for writing or speaking as these would generally have been blank responses or responses with too little content to allow rating.
It should be noted that in operational rating, as explained in Section 1, the four tasks in writing or speaking for each individual test-taker are each marked by a different rater: no rater rates more than one task for any given test-taker. However, for the purpose of the field trial, raters marked all tasks for all candidates. This was done from a practicality perspective, but also because it would allow a robustly linked data set for all tasks, test-takers, and raters in a fully crossed design for analysis with the multi-facet Rasch model software FACETS (Linacre, 2013). For both writing and speaking, a 3-facet analysis was carried out—raters, test-takers, tasks. The model for each also included separate scale definitions for each task-specific scale.

5.3.2 Writing

The FACET map for writing can be seen in Figure 2. The raters are clustered together which shows that no examiner was exceedingly harsh or lenient, with the majority of raters falling within ± 1 logits. This range of severity estimates has been suggested as a tolerable level of rater severity variability in relation to performance assessments (Van Moere, 2006; Taylor & Galaczi, 2011).The test-takers in column 3 are suitably dispersed and the tasks performed as expected with the A2 writing (W_T2) at the bottom of the logit scale in column 4, followed by the B1 task (W_T3) and then the B2 task (W_T4). Note that in the distribution of test-taker ability on the vertical logit scale, a small number of test-takers rank higher than the highest difficulty value for the tasks, which is for Task 4. This confirms the test design, in which Task 4 is targeted at a B2-level of ability with some test-takers demonstrating a level of ability beyond B2 and allocated a C-level performance.

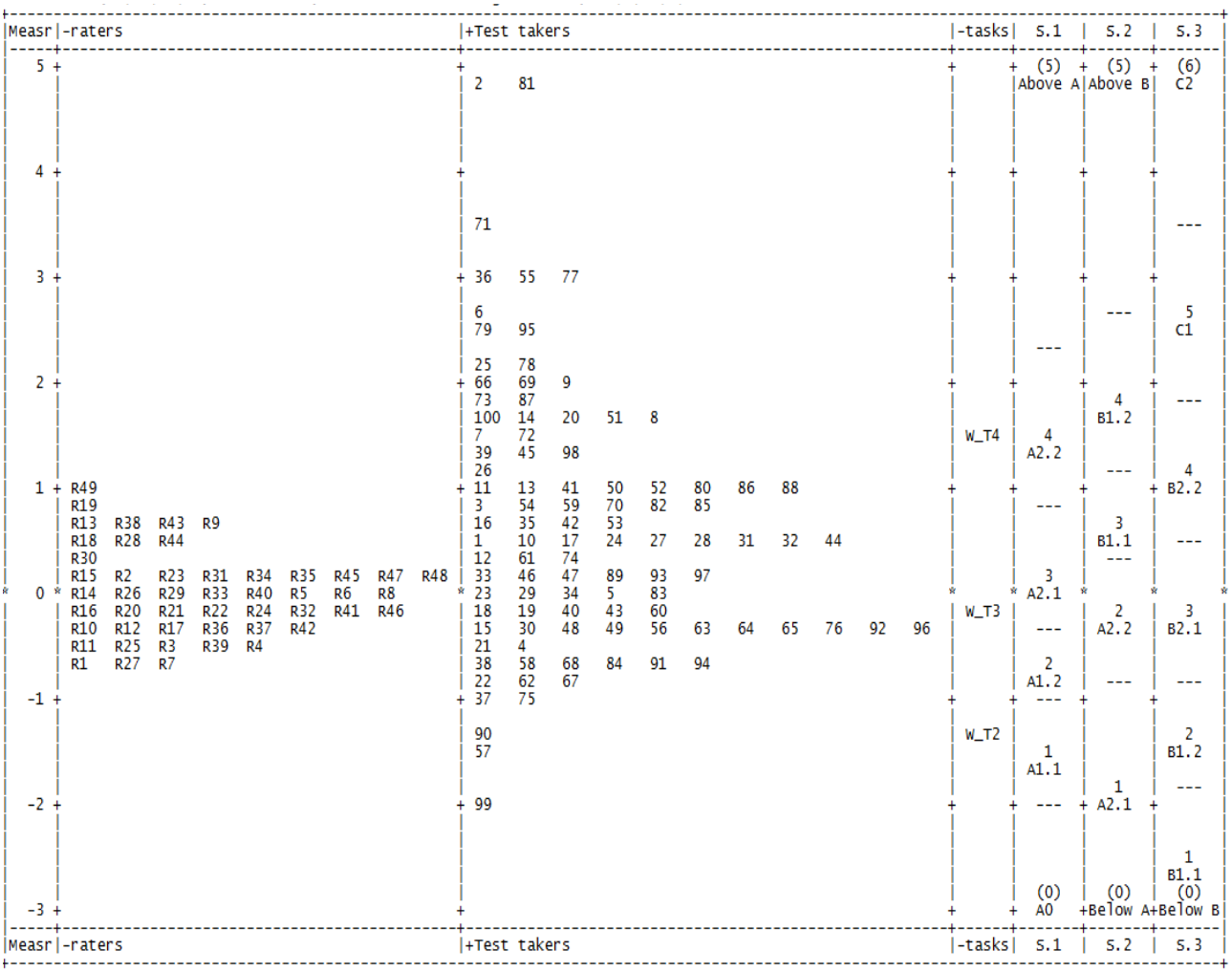
FACETS also provides measures of fit to evaluate the extent to which observed ratings provided by raters are consistent with the ratings of the other raters in the sample. The infit mean square statistics is a commonly employed indicator of rater consistency, with measures falling between 0.6–1.5 considered an acceptable range (e.g. Eckes, 2011; Engelhard & Stone, 1998; Lunz, Wright, & Linacre, 1990; O'Sullivan, 2008). Examining Appendix 6, only one rater (R25) demonstrates a result marginally outside of this range. Looking at the results for tasks themselves (Appendix 7), all tasks fall within an acceptable infit mean square range. Fit statistics are also considered to be key indicators of unidimensionality in a Rasch analysis (Bonk & Ockey, 2003; Eckes, 2009; Henning, 1992; McNamara, 1996; Sick, 2010), and the results for tasks and raters indicate that the ratings and the tasks are measuring a single construct of writing ability.

The researchers were primarily interested in the performance of the revised scales, as the tasks themselves had not changed since the test development. As such, it was expected that the tasks would perform adequately. Additionally, given the limited changes to the writing scales, and given that the raters were all trained, operational raters, a high level of consistency among raters was expected, which would be reflected through acceptable fit statistics. Adequate fit statistics for tasks and raters would thus give an indication that the revised scales for writing were being applied in a consistent and interpretable way by raters, and that the constructs embedded in the rating scales were capturing and adequately measuring a common construct of writing performance.

The last three columns show how the rating scales for the three tasks functioned. Raters used 5-point scales when rating candidates' performances on Tasks 2 and 3 (S.1 and S.2) and a 6-point scale when rating candidates' performances on Task 4 (S.3). A dotted line in a column indicates the transition point at which a candidates' probability of receiving the next higher rating for that task begins to exceed that candidate's probability of receiving the lower rating. For example, for Task 2 (S.1), the most probable rating for candidates whose writing ability measures were in the range of 2.5-5 logits was 5.

The training had been minimal before the field trial. It appears that raters were reluctant to use the highest bands (central tendency and ceiling effects). These findings were used to develop standardisation materials before going live, which emphasised that test-takers do not have to have a perfect answer to achieve a mark of 5 for the A2 and B1 tasks.

Figure 2: Facet map for writing from scale revision field trial



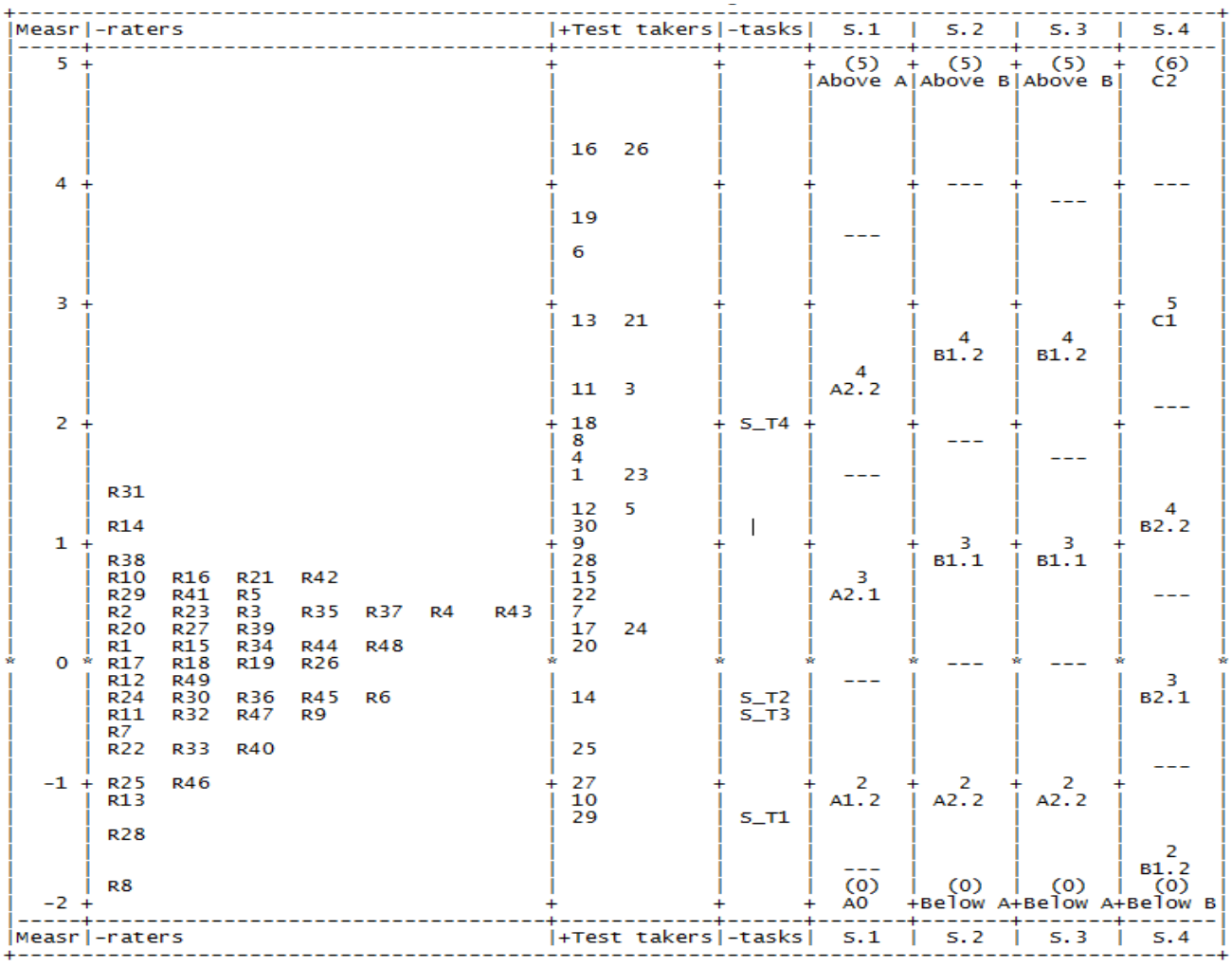
5.3.4 Speaking

Figure 3 shows the facet map for the speaking test, and as for writing, the order of difficulty for tasks is as expected with the A2 Task (S_T1) the easiest, followed by the two B1 tasks (S_T2 and S_T3) and then the B2 task (S_T4) at the top of the logit scale. For rater severity, the majority of raters fall within the +/- 1 range noted above in the discussion of writing results. However, a greater range is seen for speaking, with a small number of raters (five) falling outside that range. No particular pattern was noted however, as two raters were more severe, and three raters were more lenient than the acceptable range. To some extent, greater variation was expected for the speaking test, as the scale revision encompassed a larger degree of change than for writing, with raters having to adjust to task-specific scales rather than the single holistic scale they had been using. While greater variation was noted, the fact that the majority of raters fell within acceptable bounds was taken as an indication that the scales were interpretable and usable in a consistent way by the raters.

Turning to the rater and task measurement reports in Appendices 8 and 9, only three raters showed infit mean square statistics outside the acceptable range, indicating once again that the revised scales were being used in a consistent way by the majority of raters. As with writing, all tasks showed acceptable fit, and the results were taken as an indication that the revised scales were capturing a unidimensional construct of speaking ability and being applied in a consistent manner by raters, resulting in tasks falling in the order of difficulty as intended.

Again there appeared to be a ceiling or central tendency effect with raters reluctant to give a 5 for the A2 and B1 tasks; this was taken into consideration in the standardisation training before going live.

Figure 3: Facet map for speaking for scale revision field trial



6. DISCUSSION OF RESULTS AND CONCLUSION

The project results provided the evidence needed to give the project team sufficient confidence to recommend that the new rating scales be used operationally. In this section, the results are discussed taking into consideration the learning points collected.

The length of time to properly review, change and embed new rating scales for a global large-scale test cannot be underestimated. All stakeholders needed to be consulted at key points and communicated with in an organised way so that the changes were rolled-out in a coordinated manner. The socio-cognitive model of test development and validation provided a theoretical way to collect information from stakeholders, in particular from the raters.

Developing rating scales is not a solitary pursuit and takes at least a year to do properly. The importance of an iterative approach, collecting feedback at each stage from raters and assessment experts through questionnaires and focus groups, and conducting a small-scale and then a large-scale trial were also key to the success of the project. Constant review of the CEFR descriptors was also helpful to keep focused on the task construct and level.

There were difficult decisions to be made, such as on how much training to conduct prior to the field trial. As mentioned in the literature review, raters need to fully understand the rating scales and interpret the descriptors reliably. However, the raters were still working with the old rating scales in live rating, and there was concern that training in the revised scales before their roll-out could create confusion. It was decided that a shorter training course than would normally be used for live accrediation would be employed in preparation for the field trial, to give raters an introduction to the key changes, and that a follow-up more thorough standardisation session would be held after the field trial. This process proved to be acceptable.

The choice of type of rating scales was a key part of the project. It proved very difficult to have six or seven distinct holistic task specific levels. Trying to develop criteria for a high and low CEFR level (e.g. B1.1 and B1.2) with different criteria did not work. Using terms such as ‘mostly’, ‘frequently’, ‘occasional’ or ‘sufficient’ could be used to distinguish between the CEFR levels but not for the same high and low CEFR level. Analytic scoring was considered but would have involved a much larger change to the scoring system, both within the online platform and by raters, so was abandoned. The final product was instead a guided holistic marking system that took into consideration the test-taker’s ability to sustain their CEFR level, which it was thought would be innovative and easy to mark.

As emphasised throughout this report, the approach to test development and validation for the Aptis test system, underpinned by the socio-cognitive model, recognises the need for ongoing data collection and evaluation from multiple perspectives. While the iterative process of gathering evidence before, during and after the revision process has provided a strong body of data to support the use of the revised scales within the Aptis test system, questions still remain to be answered. Feedback from the scale revision process, as well as ongoing follow-up contact with raters by the examiner network manager suggests that raters find the scales accessible and easy to use.[1] Nonetheless, as with productive skills generally, although explicit rating scales and robust training and standardisation appear to help improve consistency and accuracy in rating, it is still not clear exactly what processes raters employ when arriving at their final decisions. As such, in addition to the ongoing collection of data regarding rater reliability which is collected and reported, it would be useful to investigate rater behaviour in more detail, in particular how the idea of sustainability is working in practice.

[1] Studies since the scales have been in place that have looked into rater reliability, including the Aptis annual operating report,

REFERENCES

Alderson, J.C. (1993). Judgements in Language Testing. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 46-57). Alexandria, VA: TESOL.

Alderson, J.C., Clapham, C. & Wall, D. (1995). *Language test construction and evaluation*. Cambridge, UK: Cambridge University Press.

Baker, B.A. (2012). Individual differences in rater decision-making style: an exploratory mixed-methods study. *Language Assessment Quarterly, 9*(3), 225–248.

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice, 18(3)*, 279–293.

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing*, *30*(4), 441–465. https://doi.org/10.1177/0265532212473244

Bejar, I.I. (2012). Rater cognition: implications for validity. *Education Measurement: Issues and Practice, 31(3),* 2–9.

British Council (2017). *Aptis Technical Update: 2015–2016*. British Council. Downloaded from https://www.britishcouncil.org/exam/aptis/research/publications/technical-report

Brown, A., Iwashita, N. & McNamara, T. (2005). An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks. *TOEFL ETS Monograph Series, 29*.

Brown, A. & Jaquith, P. (2011). The development and validation of an online rater training and marking system: promises and pitfalls. *Language Testing: Theories and Practices.* London & New York: Palgrave Macmillan.

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

Chapelle, C.M.N., Enright, M. & Jamieson, J. (2008). *Building a validity argument for the Test of English as a Foreign Language*, New York; London: Routledge.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Strasbourg, France. Online.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7(1),* 31–51.

Cumming, A., Kantor, R. & Powers, D. (2002). Decision-making while rating ESL/EFL writing tasks: a descriptive framework. *The Modern Language Journal*, 86(i), 67–96.

Dunlea, J., Fairbairn, J. & O'Sullivan, B. (2016). *Validating a scale revision project: results and implications for the underlying model of validation.* LTRC conference. Palermo, Italy.

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (Section H)*. Strasbourg, France: Council of Europe/Language Policy Division.

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

Eckes, T. (2012). Operational rater types in writing assessment: linking rater cognition to rater behaviour. *Language Assessment Quarterly, 9(3),* 270–292.

Englehard, G. & Stone, G. (1990). Evaluating the quality of ratings obtained from standard-setting judges. *Educational and Psychological Measurement*, Vol. 58 No. 2, 179–196.

Elder, C., Barkhuizen, G., Knoch, U. & Von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, 24(1), 37–64.

Elder, C., Knoch, U., Barkhuizen, G. & Von Randow, J. (2009). Individual feedback to enhance rater training: does it work?. *Language Assessment Quarterly*, 2(3), 175–196.

Fairbairn, J. (2015). *An examination of how analytic raters adapt to holistic marking*. Unpublished MA dissertation. Lancaster University.

Fulcher, G. & Davidson, F. (2007). *Language testing and assessment*. London & New York: Routledge.

Hamp-Lyons, L. (1995). Rating non-native writing: the trouble with holistic scoring. *TESOL Quarterly, 29(4),* 759–762.

Henning, G. (1992). Dimensionality and construct validity of language tests. *Language Testing, 9(1)*, 1–11.

Holsknecht, F., Eberharter, K., Kremmel, B., Zehentner, M., McCray, G., Konrad, E. & Spottl, C. (2017). *Looking into listening: Using eye-tracking to establish the cognitive validity of the Aptis Listening Test*. AR-G/2017/3. ARAGS Research Reports online. London: British Council.

Huot, B. (1990a). The literature of direct writing assessment: major concerns and prevailing trends. *Review of Educational Research, 60(2),* 237–263.

Huot, B. (1990b). Reliability, validity, and holistic scoring: what we know and what we need to know. *College Composition and Communication, 41(2),* 201–213.

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

Knoch, U., Read, J. & Von Randow, J. (2007). Re-training writing raters online: how does it compare with face-to-face training? *ScienceDirect*, *12,* 26–43.

Knoch, U. (2010). Investigating the effectiveness of individualized feedback to rating behaviour – a longitudinal study. *Language Testing, 28(2),* 179–200.

Knoch, U., Fairbairn, J. & Huisman, A. (2015). *An evaluation of the effectiveness of training Aptis raters online*. Internal British Council document available on request – journal article in progress.

Knoch, U., Fairbairn, J. & Huismann, A. (2016). An evaluation of an online rater training program for the speaking and writing sub-tests of the Aptis test. *ALTAANZ Papers in Language Testing and Assessment Special Issue, Volume 5, Issue 1.*

Lim, G.S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing, 28(4),* 543–560.

Linacre, J. (2013). *FACETS*. Computer program, version 3.71.2. Chicago: MESA Press.

Lumley, T. & McNamara, T. (1995). Rater characteristics and rater bias: implications for training. *Language Testing, 12(54),* 54–71.

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

McNamara, T. (1996). *Measuring second language performance*. London: Addison Wesley Longman.

Messick, S. (1994). The Matter of Style: Manifestations of Personality in Cognition, Learning, and Teaching. *Educational Psychologist, 29(3),* 121–136.

Myford, C.M. & Wolfe, E.W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: part 1. *Journal of Applied Measurement, 4(4),* 386–422.

O'Sullivan, B. (2008). *Modelling performance in tests of spoken language*. Bern, Switzerland: Peter Lang.

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*. TF/2015/001. London: British Council.

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

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

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

Papajohn, D. (2002). Concept Mapping for Rater Training. *TESOL Quarterly, 36(2),* 219–233.

Seedhouse, P., Harris, A., Naeb, R. & Ustunel, E. (2014). The relationship between speaking features and band descriptors: A mixed methods study. *IELTS Research Reports Online Series*. IELTS Partners: British Council, Cambridge English Language Assessment and IDP: IELTS Australia.

Shaw, S. (2002). The effect of training and standardisation on rater judgement and inter-rater reliability. *Research Notes, 8(5),* 13–17.

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

Sick, J. (2010). Assumptions and requirements of Rasch measurement. Shiken: *JALT Testing & Evaluation SIG Newsletter*. 14(2), 23–29.

Spicer, D.P. & Sadler-Smith, E. (2005). An examinations of the general decision-making style questionnaire in two UK samples. *Journal of Managerial Psychology, 20 (1/2),* 137–149.

Sweedler-Brown, C.O. (1985). The influence of training and experience on holistic essay evaluations. *The English Journal, 74(5),* 49–55.

Taylor, L. & Galaczi, E. (2011). Scoring validity. In L. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking*. Cambridge: Cambridge University Press.

Thunholm, P. (2004). Decision-making style: habit, style, or both? *Personality and Individual Differences, 36,* 931–944.

Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing*, *23*(4), 411–440. https://doi.org/10.1191/0265532206lt336oa

Watts, A. (2006). *Fostering communities of practice: A rationale for developing the use of new technologies in support of raters.* Presented at International Association for Educational Assessment conference. Cambridge, UK: Cambridge University Press.

Weigle, S.C. (1994). Effects of training on raters of ESL composition. *Language Testing, 11(2),* 197–223.

Weigle, S.C. (1998). Using FACETS to model rater training effects. *Language Testing, 15(2),* 263–287.

Weigle, S.C. (2002). *Assessing writing*. Cambridge, UK: Cambridge University Press.

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

Appendix 1: Example of an Aptis Rating scale (Speaking Tasks 2 and 3)

5 **B2 (or above)**	Likely to be above B1 level.
4 **B1.2**	Responses to all **three** questions are on topic and show the following features: • Control of simple grammatical structures. Errors occur when attempting complex structures. • Sufficient range and control of vocabulary for the task. Errors occur when expressing complex thoughts. • Pronunciation is intelligible but inappropriate mispronunciations put an occasional strain on the listener. • Some pausing, false starts and reformulations. • Uses only simple cohesive devices. Links between ideas are not always clearly indicated.
3 **B1.1**	Responses to **two** questions are on topic and show the following features: • Control of simple grammatical structures. Errors occur when attempting complex structures. • Sufficient range and control of vocabulary for the task. Errors occur when expressing complex thoughts. • Pronunciation is intelligible but inappropriate mispronunciations put an occasional strain on the listener. • Some pausing, false starts and reformulations. • Uses only simple cohesive devices. Links between ideas are not always clearly indicated.
2 **A2.2**	Responses to at least **two** questions are on topic and show the following features: • Uses some simple grammatical structures correctly but systematically makes basic mistakes. • Vocabulary will be limited to concrete topics and descriptions. Inappropriate lexical choices for the task are noticeable. • Mispronunciations are noticeable and put a strain on the listener. • Noticeable pausing, false starts and reformulations. • Cohesion between ideas is limited. Responses tend to be a list of points.
1 **A2.1**	Response to **one** question is on topic and shows the following features: • Uses some simple grammatical structures correctly but systematically makes basic mistakes. • Vocabulary will be limited to concrete topics and descriptions. Inappropriate lexical choices for the task are noticeable. • Mispronunciations are noticeable and put a strain on the listener. • Noticeable pausing, false starts and reformulations. • Cohesion between ideas is limited. Responses tend to be a list of points.
0	• Performance below A2.

Appendix 2: Questionnaire for raters

Indicate to what extent you agree with the following statements when marking Writing Tasks.

1. The rating scale is useful for evaluating performances on Writing Task 2.
   (Strongly disagree / Disagree / Agree / Strongly agree)
2. The descriptions of performance for each Writing Task 2 score band are clear.
   (Strongly disagree / Disagree / Agree / Strongly agree)
3. The distinctions between different Writing Task 2 score bands are clear.
   (Strongly disagree / Disagree / Agree / Strongly agree)
4. The number of points on the scale (0-5) is appropriate for rating Writing Task 2.
   (Strongly disagree / Disagree / Agree / Strongly agree)
5. I would prefer to give separate scores for each element of performance in Writing Task 2 (e.g. separate scores for grammar, vocabulary, etc.).
   (Strongly disagree / Disagree / Agree / Strongly agree)
6. Indicate how much importance you place on the following when rating Writing Task 2.
   (None / A little / Some / A lot)
   - Coherence (clarity and organisation of message)
   - Cohesion (use of connecting devices and discourse markers)
   - Grammatical accuracy
   - Grammatical range
   - Overall impression
   - Punctuation
   - Register
   - Relevance of content to the topic
   - Spelling
   - Task completion
   - Vocabulary accuracy
   - Vocabulary range

Questions 1 – 6 were repeated for Tasks 3 and 4.

7. The rating scale is useful for evaluating performances on SPEAKING TASK 1.
   (Strongly disagree / Disagree / Agree / Strongly agree)
8. The rating scale is useful for evaluating performances on SPEAKING TASK 2.
   (Strongly disagree / Disagree / Agree / Strongly agree)
9. The rating scale is useful for evaluating performances on SPEAKING TASK 3.
   (Strongly disagree / Disagree / Agree / Strongly agree)
10. The rating scale is useful for evaluating performances on SPEAKING TASK 4.
    (Strongly disagree / Disagree / Agree / Strongly agree)
11. The wording used to describe performance in each SPEAKING score band is clear.
    (Strongly disagree / Disagree / Agree / Strongly agree)
12. The distinctions between different SPEAKING score bands are clear.
    (Strongly disagree / Disagree / Agree / Strongly agree)
13. The number of points on the scale (0-5) is appropriate for rating SPEAKING tasks.
    (Strongly disagree / Disagree / Agree / Strongly agree)
14. I would prefer to give separate scores for each element of SPEAKING performance (e.g. separate scores for grammar, vocabulary, etc).
    (Strongly disagree / Disagree / Agree / Strongly agree)

15. Indicate how much importance you place on the following when rating Speaking tasks.
    (None / A little / Some / A lot)
    - Coherence (clarity and organisation of message)
    - Cohesion (use of connecting devices and discourse markers)
    - Fluency (speech rate and pausing)
    - Grammatical accuracy
    - Grammatical range
    - Intonation
    - Overall impression
    - Pronunciation
    - Relevance of content to the topic
    - Task completion
    - Vocabulary accuracy
    - Vocabulary range
16. I read the task and task instructions before rating the performance on a task.
    (Never / Sometimes / Usually / Always)
17. I read/listen to the entire performance for a task before I think about a rating.
    (Never / Sometimes / Usually / Always)
18. I refer to the descriptions of performance in the rating scales when deciding on a final rating for a task.
    (Never / Sometimes / Usually / Always)
19. I rely mainly on my first overall impression when deciding on a final task rating.
    (Never / Sometimes / Usually / Always)
20. I consider each of the aspects of performance described in a score band before deciding on a final rating for a task.
    (Never / Sometimes / Usually / Always)
21. I think of a score for each of the aspects described in the scale (e.g. a 4 for grammar, a 5 for vocabulary) and average those scores to derive a final task rating.
    (Never / Sometimes / Usually / Always)
22. Comments

Appendix 3: Rater questionnaire results

A2 Writing task scale (Task 2)

Writing A2 Task	Strongly Disagree	Disagree	Agree	Strongly Agree
The rating scale is useful for evaluating performances on Writing Task 2.	1	4	25	8
The descriptions of performance for each Writing Task 2 score band are clear.	1	4	24	9
The distinctions between different Writing Task 2 score bands are clear.	1	6	24	7
The number of points on the scale (0 – 5) is appropriate for rating Writing Task 2.	1	1	28	8
I would prefer to give separate scores for each element of performance in Writing Task 2 (e.g. separate scores for grammar, vocabulary, etc.).	9	21	7	1

B1 Writing task scale (Task 3)

Writing A2 Task	Strongly Disagree	Disagree	Agree	Strongly Agree
The rating scale is useful for evaluating performances on Writing Task 3.	1	10	7	10
The descriptions of performance for each Writing Task 3 score band are clear.	1	0	30	7
The distinctions between different Writing Task 3 score bands are clear.	1	4	28	5
The number of points on the scale (0 – 5) is appropriate for rating Writing Task 3.	1	0	25	12
I would prefer to give separate scores for each element of performance in Writing Task 3 (e.g. separate scores for grammar, vocabulary, etc.).	9	20	7	2

B2 Writing task scale (Task 4)

Writing A2 Task	Strongly Disagree	Disagree	Agree	Strongly Agree
The rating scale is useful for evaluating performances on Writing Task 4.	1	2	27	8
The descriptions of performance for each Writing Task 4 score band are clear.	1	4	27	6
The distinctions between different Writing Task 4 score bands are clear.	1	5	25	7
The number of points on the scale (0 – 5) is appropriate for rating Writing Task 4.	0	1	24	13
I would prefer to give separate scores for each element of performance in Writing Task 4 (e.g. separate scores for grammar, vocabulary, etc.).	8	18	8	12

Speaking scale used for all four tasks

	Strongly Disagree	Disagree	Agree	Strongly Agree
The rating scale is useful for evaluating performances on Speaking Task 1.	1	0	23	13
The rating scale is useful for evaluating performances on Speaking Task 2.	1	0	25	11
The rating scale is useful for evaluating performances on Speaking Task 3.	1	0	25	11
The rating scale is useful for evaluating performances on Speaking Task 4.	1	2	23	11
The wording used to describe performance in each Speaking score band is clear.	1	2	26	8
The distinctions between different Speaking score bands are clear.	3	2	25	7
The number of points on the scale (0 – 5) is appropriate for rating Speaking.	2	2	21	12
I would prefer to give separate scores for each element of performance in Speaking (e.g. separate scores for grammar, vocabulary, etc.).	9	16	9	3

Rater decision-making

	Never	Sometimes	Usually	Always
I read the task and task instructions before rating the performance on a task.	0	3	10	24
I read/listen to the entire performance for a task before I think about a rating.	1	8	11	17
I refer to the descriptions of performance in the rating scales when deciding on a final rating for a task.	0	9	13	15
I rely mainly on my first overall impression when deciding on a final task rating.	9	20	8	0
I consider each of the aspects of performance described in a score band before deciding on a final rating for a task.	0	8	15	14
I think of a score for each of the aspects described in the scale (e.g. a 4 for grammar, a 5 for vocabulary) and average those scores to derive a final task rating.	19	12	5	1

Appendix 4: Questionnaire for small-scale pilot

1. The rating scale is useful for evaluating performances on Writing Task 2.
   (Strongly disagree / Disagree / Agree / Strongly agree)
2. The descriptions of performance for each Writing Task 2 score band are clear.
   (Strongly disagree / Disagree / Agree / Strongly agree)
3. The distinctions between different Writing Task 2 score bands are clear.
   (Strongly disagree / Disagree / Agree / Strongly agree)
4. The number of points on the scale (0 – 5) is appropriate for rating Writing Task 2.
   (Strongly disagree / Disagree / Agree / Strongly agree)
5. How would a text that is too long impact on your marking of Writing Task 2?
   (for Speaking rating scales this question is changed to: How would a partially inaudible item impact on marking?)
6. Do you think the new Writing Task 2 scale allows you to mark more accurately?
   (Yes / No)
7. What did you least like about the new Writing Task 2 rating scale?
   Any difficulties or problems?
8. What did you most like about the new Writing Task 2 rating scale?

Questions 1 – 8 were repeated for all rating scales.

9. Which rating scale did you find most difficult to use?
10. Which rating scale did you easiest to use?
11. Please give your overall view of the updated rating scales and any additional feedback you think is useful to help us improve the rating scales.

Appendix 5: Small-scale pilot questionnaire results

A2 Writing task scale (Task 2)

	Strongly Disagree	Disagree	Agree	Strongly Agree
The rating scale is useful for evaluating performances on Writing Task 2.	0	0	3	4
The descriptions of performance for each Writing Task 2 score band are clear.	0	0	3	4
The distinctions between different Writing Task 2 score bands are clear.	0	0	4	2
The number of points on the scale (0 – 5) is appropriate for rating Writing Task 2.	0	0	2	5

B1 Writing task scale (Task 3)

	Strongly Disagree	Disagree	Agree	Strongly Agree
The rating scale is useful for evaluating performances on Writing Task 3.	0	0	5	2
The descriptions of performance for each Writing Task 3 score band are clear.	0	2	4	1
The distinctions between different Writing Task 3 score bands are clear.	0	2	4	1
The number of points on the scale (0 – 5) is appropriate for rating Writing Task 3.	0	0	3	4

B2 Writing task scale (Task 4)

	Strongly Disagree	Disagree	Agree	Strongly Agree
The rating scale is useful for evaluating performances on Writing Task 4.	0	0	3	3
The descriptions of performance for each Writing Task 4 score band are clear.	0	1	4	0
The distinctions between different Writing Task 4 score bands are clear.	0	1	4	1
The number of points on the scale (0 – 6) is appropriate for rating Writing Task 4.	0	0	3	3

A2 Speaking task scale (Task 1)

	Strongly Disagree	Disagree	Agree	Strongly Agree
The rating scale is useful for evaluating performances on Speaking Task 1.	0	0	4	2
The descriptions of performance for each Speaking Task 1 score band are clear.	0	1	3	2
The distinctions between different Speaking Task 1 score bands are clear.	0	1	3	2
The number of points on the scale (0 – 5) is appropriate for rating Speaking Task 1.	0	1	2	3

B1 Speaking task scale (Task 2)

	Strongly Disagree	Disagree	Agree	Strongly Agree
The rating scale is useful for evaluating performances on Speaking Task 2.	0	1	4	2
The descriptions of performance for each Speaking Task 2 score band are clear.	0	2	3	2
The distinctions between different Speaking Task 2 score bands are clear.	0	1	2	3
The number of points on the scale (0 – 5) is appropriate for rating Speaking Task 2.	0	1	4	2

B1 Speaking task scale (Task 3)

	Strongly Disagree	Disagree	Agree	Strongly Agree
The rating scale is useful for evaluating performances on Speaking Task 3.	0	2	4	1
The descriptions of performance for each Speaking Task 3 score band are clear.	0	3	3	1
The distinctions between different Speaking Task 3 score bands are clear.	0	3	3	1
The number of points on the scale (0 – 5) is appropriate for rating Speaking Task 3.	0	2	4	1

B2 Speaking task scale (Task 4)

	Strongly Disagree	Disagree	Agree	Strongly Agree
The rating scale is useful for evaluating performances on Speaking Task 4.	0	0	7	0
The descriptions of performance for each Speaking Task 4 score band are clear.	0	2	5	0
The distinctions between different Speaking Task 4 score bands are clear.	0	2	5	0
The number of points on the scale (0 – 6) is appropriate for rating Speaking Task 4.	0	0	4	2

Appendix 6: Rater score and measure file for writing

1	raters																			
T.Score	T.Count	Obs.Avge	FairMAvge	Measure	S.E.	InfitMS	InfitZ	OutfitMS	OutfitZ	PtBis	PtMeExp	Discrim	Displace	Status	Group	Weight	1	raters	F-Number	F-Label
1100	300	3.67	3.88	-0.72	0.07	0.95	-0.6	1.11	0.95	0.52	0.78	1.02	0	-1	0	1	1	R1	1	raters
898	299	3	3.09	0.21	0.07	0.78	-2.9	0.83	-1.93	0.56	0.8	1.22	0	-1	0	1	2	R2	1	raters
1048	300	3.49	3.68	-0.47	0.07	0.95	-0.53	0.9	-0.91	0.55	0.78	1.15	0	-1	0	1	3	R3	1	raters
1036	299	3.46	3.64	-0.42	0.07	0.84	-2.02	0.93	-0.58	0.53	0.78	1.09	0	-1	0	1	4	R4	1	raters
958	299	3.2	3.33	-0.06	0.07	1.03	0.36	0.97	-0.22	0.59	0.79	1.11	0	-1	0	1	5	R5	1	raters
949	299	3.17	3.3	-0.02	0.07	1.1	1.24	1.06	0.68	0.57	0.79	1.02	0	-1	0	1	6	R6	1	raters
1075	297	3.62	3.81	-0.63	0.07	0.96	-0.39	1.11	0.92	0.49	0.78	0.95	0	-1	0	1	7	R7	1	raters
960	299	3.21	3.34	-0.07	0.07	0.87	-1.62	0.84	-1.65	0.58	0.79	1.2	0	-1	0	1	8	R8	1	raters
792	296	2.68	2.71	0.65	0.07	0.95	-0.57	0.92	-0.94	0.6	0.81	1.04	0	-1	0	1	9	R9	1	raters
1027	300	3.42	3.59	-0.37	0.07	1.25	2.75	1.33	2.84	0.49	0.79	0.76	0	-1	0	1	10	R10	1	raters
1045	298	3.51	3.68	-0.48	0.07	1.14	1.63	1.15	1.33	0.51	0.78	0.89	0	-1	0	1	11	R11	1	raters
1013	299	3.39	3.55	-0.31	0.07	0.78	-2.78	0.76	-2.46	0.55	0.79	1.17	0	-1	0	1	12	R12	1	raters
801	300	2.67	2.71	0.64	0.07	1	0.03	0.97	-0.32	0.58	0.81	1.02	0	-1	0	1	13	R13	1	raters
937	300	3.12	3.24	0.04	0.07	0.78	-2.87	0.79	-2.26	0.57	0.8	1.25	0	-1	0	1	14	R14	1	raters
873	291	3	3.09	0.22	0.07	0.92	-0.94	0.9	-1.05	0.56	0.8	1.06	0	-1	0	1	15	R15	1	raters
982	299	3.28	3.42	-0.17	0.07	0.77	-3.01	0.81	-1.97	0.56	0.79	1.21	0	-1	0	1	16	R16	1	raters
1032	300	3.44	3.61	-0.39	0.07	0.95	-0.54	0.93	-0.63	0.52	0.78	1.01	0	-1	0	1	17	R17	1	raters
819	300	2.73	2.78	0.56	0.07	0.96	-0.49	1.03	0.35	0.57	0.81	1.09	0	-1	0	1	18	R18	1	raters
755	295	2.56	2.59	0.77	0.07	0.8	-2.55	0.8	-2.42	0.59	0.8	1.19	0	-1	0	1	19	R19	1	raters
964	299	3.22	3.35	-0.08	0.07	0.9	-1.29	0.84	-1.65	0.58	0.79	1.19	0	-1	0	1	20	R20	1	raters
990	299	3.31	3.46	-0.21	0.07	1.12	1.37	1.03	0.37	0.57	0.79	1	0	-1	0	1	21	R21	1	raters
977	300	3.26	3.4	-0.14	0.07	1.48	5.08	1.59	5.02	0.45	0.79	0.44	0	-1	0	1	22	R22	1	raters
906	299	3.03	3.12	0.17	0.07	1.32	3.58	1.36	3.42	0.52	0.8	0.72	0	-1	0	1	23	R23	1	raters
982	299	3.28	3.42	-0.17	0.07	1.07	0.8	1	0.05	0.56	0.79	1.03	0	-1	0	1	24	R24	1	raters
1057	297	3.56	3.74	-0.54	0.07	1.55	5.56	1.95	6.64	0.4	0.78	0.3	0	-1	0	1	25	R25	1	raters
951	300	3.17	3.29	-0.02	0.07	0.83	-2.22	0.83	-1.75	0.57	0.79	1.18	0	-1	0	1	26	R26	1	raters
1070	293	3.65	3.83	-0.65	0.07	0.98	-0.25	0.92	-0.64	0.51	0.77	1.08	0	-1	0	1	27	R27	1	raters
824	299	2.76	2.8	0.53	0.07	0.73	-3.64	0.77	-2.76	0.59	0.8	1.2	0	-1	0	1	28	R28	1	raters
959	297	3.23	3.34	-0.08	0.07	1	-0.02	0.99	-0.02	0.56	0.79	1.06	0	-1	0	1	29	R29	1	raters
886	300	2.95	3.04	0.27	0.07	0.77	-3	0.83	-1.95	0.56	0.8	1.17	0	-1	0	1	30	R30	1	raters
922	299	3.08	3.2	0.09	0.07	1.05	0.63	1.17	1.7	0.5	0.8	0.83	0	-1	0	1	31	R31	1	raters
977	299	3.27	3.41	-0.15	0.07	0.83	-2.18	0.82	-1.88	0.56	0.79	1.17	0	-1	0	1	32	R32	1	raters
956	300	3.19	3.31	-0.04	0.07	1.19	2.19	1.26	2.44	0.51	0.79	0.73	0	-1	0	1	33	R33	1	raters
900	299	3.01	3.1	0.2	0.07	1.06	0.72	1	0.07	0.58	0.8	1.01	0	-1	0	1	34	R34	1	raters
902	298	3.03	3.12	0.18	0.07	1.13	1.56	1.17	1.74	0.54	0.8	0.84	0	-1	0	1	35	R35	1	raters
1019	300	3.4	3.56	-0.33	0.07	0.85	-1.82	0.82	-1.75	0.55	0.79	1.16	0	-1	0	1	36	R36	1	raters
1037	300	3.46	3.63	-0.41	0.07	0.88	-1.41	0.86	-1.28	0.53	0.78	1.11	0	-1	0	1	37	R37	1	raters
775	288	2.69	2.68	0.67	0.07	0.85	-1.87	1.03	0.36	0.55	0.8	1.03	0	-1	0	1	38	R38	1	raters
1049	299	3.51	3.69	-0.48	0.07	1.21	2.38	1.37	3.01	0.48	0.78	0.76	0	-1	0	1	39	R39	1	raters
929	299	3.11	3.21	0.07	0.07	0.79	-2.76	1.27	2.61	0.53	0.8	1.05	0	-1	0	1	40	R40	1	raters
968	300	3.23	3.36	-0.09	0.07	0.75	-3.37	1.08	0.82	0.55	0.79	1.16	0	-1	0	1	41	R41	1	raters
1029	300	3.43	3.6	-0.38	0.07	0.97	-0.34	0.99	-0.08	0.55	0.79	1.07	0	-1	0	1	42	R42	1	raters
779	300	2.6	2.63	0.73	0.07	0.99	-0.09	0.96	-0.44	0.58	0.81	1.09	0	-1	0	1	43	R43	1	raters
846	299	2.83	2.89	0.44	0.07	1.02	0.23	0.97	-0.31	0.58	0.8	1.01	0	-1	0	1	44	R44	1	raters
919	299	3.07	3.17	0.12	0.07	0.84	-2.01	0.94	-0.62	0.53	0.8	1.02	0	-1	0	1	45	R45	1	raters
957	297	3.22	3.36	-0.09	0.07	1.3	3.31	1.44	3.86	0.47	0.79	0.56	0	-1	0	1	46	R46	1	raters
896	300	2.99	3.08	0.23	0.07	1.01	0.17	1.12	1.25	0.51	0.8	0.82	0	-1	0	1	47	R47	1	raters
911	300	3.04	3.14	0.16	0.07	1.04	0.48	1.2	1.99	0.49	0.8	0.77	0	-1	0	1	48	R48	1	raters
716	300	2.39	2.38	1.01	0.07	0.96	-0.49	0.94	-0.64	0.59	0.81	1.05	0	-1	0	1	49	R49	1	raters

Appendix 7: Task score and measure file for writing

3	tasks																			
T.Score	T.Count	Obs.Avge	FairMAvge	Measure	S.E.	InfitMS	InfitZ	OutfitMS	OutfitZ	PtBis	PtMeExp	Discrim	Displace	Status	Group	Weight	3	tasks	F-Number	F-Label
20000	4898	4.08	4.28	-1.31	0.02	1.1	4.14	1.16	5.05	0.41	0.6	0.89	0	-1	0	1	1	W_T2	3	tasks
14374	4880	2.95	3.06	-0.13	0.02	0.85	-8.17	0.89	-5.08	0.55	0.72	1.17	0	-1	0	1	2	W_T3	3	tasks
11779	4850	2.43	2.49	1.44	0.02	1.05	2.33	1.05	2.25	0.51	0.75	0.95	0	-1	0	1	3	W_T4	3	tasks

Appendix 8: Rater score and measure file for speaking

1 raters																				
T.Score	T.Count	Obs.Avge	FairMAvge	Measure	S.E.	InfitMS	InfitZ	OutfitMS	OutfitZ	PtBis	PtMeExp	Discrim	Displace	Status	Group	Weight	1	raters	F-Number	F-Label
375	116	3.23	3.35	0.18	0.13	1.35	2.45	1.53	3.43	0.39	0.83	0.51	0	-1	0	1	1	R1	1	raters
332	108	3.07	3.16	0.47	0.13	2.29	7.04	2.8	8.95	0.21	0.83	-0.52	0	-1	0	1	2	R2	1	raters
355	116	3.06	3.16	0.49	0.12	0.69	-2.6	0.68	-2.66	0.63	0.83	1.37	0	-1	0	1	3	R3	1	raters
356	114	3.12	3.21	0.41	0.13	0.96	-0.24	1.02	0.16	0.56	0.83	1.05	0	-1	0	1	4	R4	1	raters
351	116	3.03	3.12	0.55	0.12	0.97	-0.17	0.99	0	0.62	0.83	1.04	0	-1	0	1	5	R5	1	raters
402	116	3.47	3.6	-0.25	0.13	0.92	-0.57	0.86	-0.98	0.61	0.82	1.2	0	-1	0	1	6	R6	1	raters
415	114	3.64	3.79	-0.58	0.13	0.95	-0.34	0.95	-0.28	0.53	0.81	1.09	0	-1	0	1	7	R7	1	raters
477	112	4.26	4.48	-1.89	0.15	0.64	-2.74	0.53	-2.5	0.56	0.76	1.38	0	-1	0	1	8	R8	1	raters
408	114	3.58	3.72	-0.46	0.13	0.58	-3.7	0.53	-3.87	0.61	0.81	1.47	0	-1	0	1	9	R9	1	raters
343	115	2.98	3.06	0.65	0.12	0.61	-3.4	0.64	-3.07	0.6	0.83	1.37	0	-1	0	1	10	R10	1	raters
415	116	3.58	3.73	-0.46	0.13	0.77	-1.87	0.76	-1.81	0.58	0.81	1.27	0	-1	0	1	11	R11	1	raters
383	113	3.39	3.52	-0.12	0.13	0.92	-0.58	0.99	0	0.52	0.82	1.03	0	-1	0	1	12	R12	1	raters
453	116	3.91	4.1	-1.12	0.13	0.94	-0.43	0.82	-1.11	0.55	0.79	1.16	0	-1	0	1	13	R13	1	raters
313	115	2.72	2.81	1.08	0.12	1.22	1.61	1.26	1.86	0.54	0.83	0.74	0	-1	0	1	14	R14	1	raters
374	116	3.22	3.34	0.2	0.13	0.89	-0.82	0.9	-0.74	0.58	0.83	1.15	0	-1	0	1	15	R15	1	raters
320	108	2.96	3.03	0.71	0.13	1	0.06	1	0.05	0.58	0.84	0.98	0	-1	0	1	16	R16	1	raters
381	114	3.34	3.49	-0.05	0.13	0.68	-2.62	0.7	-2.4	0.59	0.82	1.29	0	-1	0	1	17	R17	1	raters
384	116	3.31	3.43	0.04	0.13	0.61	-3.4	0.67	-2.68	0.56	0.82	1.29	0	-1	0	1	18	R18	1	raters
383	115	3.33	3.46	-0.01	0.13	0.95	-0.32	0.95	-0.33	0.58	0.82	1.08	0	-1	0	1	19	R19	1	raters
367	115	3.19	3.28	0.27	0.13	1.05	0.42	1.02	0.18	0.59	0.83	1	0	-1	0	1	20	R20	1	raters
342	115	2.97	3.07	0.65	0.12	1.31	2.21	1.25	1.79	0.57	0.83	0.78	0	-1	0	1	21	R21	1	raters
432	116	3.72	3.89	-0.75	0.13	0.67	-2.71	0.73	-1.88	0.56	0.8	1.25	0	-1	0	1	22	R22	1	raters
347	111	3.13	3.21	0.4	0.13	1.21	1.5	1.12	0.86	0.58	0.83	0.89	0	-1	0	1	23	R23	1	raters
402	116	3.47	3.6	-0.25	0.13	0.97	-0.17	0.95	-0.35	0.54	0.82	0.99	0	-1	0	1	24	R24	1	raters
433	113	3.83	4	-0.94	0.13	1.27	1.83	1.29	1.72	0.5	0.8	0.68	0	-1	0	1	25	R25	1	raters
384	115	3.34	3.45	0	0.13	1.17	1.24	1.13	0.96	0.55	0.82	0.8	0	-1	0	1	26	R26	1	raters
363	114	3.18	3.29	0.27	0.13	1.12	0.92	1.09	0.7	0.57	0.83	0.89	0	-1	0	1	27	R27	1	raters
463	115	4.03	4.25	-1.4	0.14	0.74	-2.02	0.85	-0.8	0.52	0.78	1.18	0	-1	0	1	28	R28	1	raters
349	116	3.01	3.1	0.58	0.12	0.7	-2.54	0.66	-2.85	0.6	0.83	1.31	0	-1	0	1	29	R29	1	raters
396	114	3.47	3.61	-0.27	0.13	0.88	-0.93	0.86	-0.96	0.58	0.82	1.13	0	-1	0	1	30	R30	1	raters
282	112	2.52	2.58	1.48	0.13	1.13	0.96	1.1	0.79	0.59	0.84	0.9	0	-1	0	1	31	R31	1	raters
412	116	3.55	3.7	-0.41	0.13	0.93	-0.51	1.04	0.3	0.47	0.81	0.92	0	-1	0	1	32	R32	1	raters
429	115	3.73	3.89	-0.75	0.13	0.79	-1.61	0.68	-2.28	0.6	0.8	1.34	0	-1	0	1	33	R33	1	raters
372	115	3.23	3.33	0.2	0.13	0.94	-0.42	0.94	-0.44	0.56	0.82	1.07	0	-1	0	1	34	R34	1	raters
351	113	3.11	3.19	0.44	0.13	0.76	-1.95	0.73	-2.2	0.63	0.83	1.35	0	-1	0	1	35	R35	1	raters
396	113	3.5	3.65	-0.32	0.13	0.67	-2.77	0.82	-1.3	0.58	0.82	1.28	0	-1	0	1	36	R36	1	raters
359	116	3.09	3.2	0.43	0.12	1.81	5.02	1.72	4.53	0.42	0.83	0.14	0	-1	0	1	37	R37	1	raters
330	114	2.89	2.98	0.79	0.13	1.11	0.82	1.24	1.72	0.56	0.83	0.83	0	-1	0	1	38	R38	1	raters
355	112	3.17	3.3	0.27	0.13	0.67	-2.78	0.67	-2.74	0.61	0.83	1.35	0	-1	0	1	39	R39	1	raters
424	115	3.69	3.84	-0.66	0.13	1.15	1.08	1.28	1.79	0.5	0.81	0.79	0	-1	0	1	40	R40	1	raters
347	114	3.04	3.15	0.51	0.13	1.09	0.71	1.14	1.04	0.53	0.83	0.84	0	-1	0	1	41	R41	1	raters
343	115	2.98	3.06	0.65	0.12	1.07	0.57	1.09	0.68	0.49	0.83	0.86	0	-1	0	1	42	R42	1	raters
360	116	3.1	3.21	0.41	0.12	0.88	-0.88	0.87	-0.95	0.58	0.83	1.18	0	-1	0	1	43	R43	1	raters
367	113	3.25	3.36	0.17	0.13	1.37	2.53	1.37	2.51	0.57	0.83	0.67	0	-1	0	1	44	R44	1	raters
404	116	3.48	3.62	-0.28	0.13	1.13	1.01	1.02	0.15	0.59	0.82	1.05	0	-1	0	1	45	R45	1	raters
444	115	3.86	4.05	-1.03	0.13	0.67	-2.7	0.74	-1.72	0.56	0.8	1.29	0	-1	0	1	46	R46	1	raters
409	116	3.53	3.67	-0.37	0.13	1.27	1.91	1.22	1.48	0.51	0.81	0.76	0	-1	0	1	47	R47	1	raters
372	115	3.23	3.34	0.18	0.13	1.13	1.02	1.22	1.56	0.47	0.83	0.78	0	-1	0	1	48	R48	1	raters
389	114	3.41	3.5	-0.09	0.13	1.28	1.95	1.16	1.12	0.55	0.81	0.88	0	-1	0	1	49	R49	1	raters

Appendix 9: Task score and measure file for speaking

3 tasks																				
T.Score	T.Count	Obs.Avge	FairMAvge	Measure	S.E.	InfitMS	InfitZ	OutfitMS	OutfitZ	PtBis	PtMeExp	Discrim	Displace	Status	Group	Weight	3	tasks	F-Number	F-Label
5532	1393	3.97	4.12	-1.34	0.04	1.19	4.52	1.19	3.76	0.47	0.71	0.8	0	-1	0	1	1	S_T1	3	tasks
4728	1414	3.34	3.39	-0.31	0.04	0.89	-3.08	0.89	-2.91	0.55	0.75	1.13	0	-1	0	1	2	S_T2	3	tasks
4757	1402	3.39	3.44	-0.36	0.04	0.86	-3.94	0.89	-3.09	0.55	0.75	1.13	0	-1	0	1	3	S_T3	3	tasks
3626	1396	2.6	2.64	2.01	0.03	1.07	1.74	1.07	1.86	0.55	0.81	0.93	0	-1	0	1	4	S_T4	3	tasks

BRITISH COUNCIL
APTIS TECHNICAL REPORTS

Aptis Speaking and Writing Rating Scales Revision Technical Report

Judith Fairbairn, British Council
Jamie Dunlea, British Council

www.britishcouncil.org/aptis

ISSN 2057-7168