City & Guilds Communicator IESOL Examination (B2) CEFR Linking Project

Case Study Report

Prepared by

Professor Barry O'Sullivan, Roehampton University, London

October 2008

Acknowledgements

I would like to thank the following organisations and individuals who have contributed to this project:

My primary thanks are due to City & Guilds of London for initiating and funding this project. In particularly I would like to thank Rachel Roberts, who contributed greatly to the project and to this report.

I would also like to thank the participating expert panel members, the teaching centres and learners who contributed so much to the project.

Finally, I would like to thank Elif Kantargioglu from Bilkent University, Ankara who acted as an independent reviewer of this final report. Elif's valuable and constructive criticism contributed to the reports' final form.

Table of Contents

EXECUTIVE SUMMARY I

PART 1 – BACKGROUND TO THE PROJECT 4

1.1. The Purpose of the Project 5

1.2. The Communicator Test 6
1.2.1. Structure of the Test 6
1.2.2. Intended & Actual Audience 7

PART 3 – THE FAMILIARISATION STAGE 10

3.1. Development phase 10

3.2. Ongoing familiarisation 10

3.3. Iterative familiarisation 13

PART 4 – THE SPECIFICATION STAGE 15

4.1. Preliminary Decisions 15

4.2. Forms completed 18

4.3. Procedure 18

4.4. Lessons Learnt 19

4.4. Some Observations on this Process 20

4.5. Evidence and Claims from this Process 21

PART 5 – THE STANDARDISATION STAGE 23

5.1. The Methodology 24

5.2. The Expert Panel (Critical Review) 25
5.2.1. Objectives 25
5.2.2. Composition 26
5.2.3. Procedure 27
5.2.4. Outcomes from the Preliminary Expert Panel 28
5.2.5. Task Re-Specification & Trials 31
5.2.6. Analysis of the Trial Data (Listening) 32
5.2.7. Analysis of the Trial Data (Reading) 39
5.2.8. Analysis of the Trial Data (teacher & self-assessment) 44

5.3. The Expert Panel (Standard Setting) 46
5.3.1. The First Panel Event (Reading) 47
5.3.2. The Second Panel Event (Listening) 51
5.3.3. The Third Panel Event (Writing) 54

5.4. Claims 61

PART 6 – THE VALIDATION STAGE 63

6.1. The Test Taker 63

6.2. Context Validity Evidence 64

6.3. Scoring Validity Evidence 68

6.4. Criterion-Related Validity Evidence 69
6.4.1. The Criterion Study: The Reading Paper 70
6.4.2. The Criterion Study: The Listening Paper 73
6.4.3. The Criterion Study: The Writing Paper 79

6.5. Claims 80

PART 7 – SUMMARY AND DISCUSSION 82

7.1. Summary of the Project 82

7.2. Summary of the Main Findings and Claims 84

7.3. Implications of the Project 85
7.3.1. For the Communicator Examination 85
7.3.2. For City & Guilds and Other Examination Boards 86
7.3.3. For the CEFR Linking Manual 86
7.3.4. For the CEFR Itself 90
7.3.5. Limitations 90

7.4. Concluding Comments 91

8. REFERENCES 92

9. APPENDICES 94

Appendix 1 Communicator level – B2 Test Syllabus with CEFR Linking Rationale 95

Appendix 2 Completed Specification Forms 104

Appendix 3 Extracts from FACETS output (Preliminary Expert Panel) 137

Appendix 4 Self Assessment 'Can DO' Instrument 143

Appendix 5 Sample Communicator Paper (CEFR B2) 144

Appendix 6 Example of Task Specific Scale (Task 1) 162

Executive Summary

Background

This project was a joint undertaking by City & Guilds and the Centre of Language Assessment Research (CLARe) at Roehampton University. The object of the project was to provide evidence of the validity of City & Guilds' Communicator examination, particularly in relation to the central claim that it is aimed at Level B2 in the Common European Framework of Reference for Languages (commonly referred to as the CEFR). In doing this, it was planned that the project would act as a formal review of the existing examination, and it was planned that any areas of concern within the papers would be identified and brought into line with best practice in the area.

The Communicator (and the other examinations in the suite) was developed using the CEFR (Council of Europe 2001) as source document to inform the assessment tasks, specifications and assessment criteria. During the development phase, however, the Draft Manual (2003) for relating examinations to the framework was not in existence, so the organisation embarked on a series of internal activities to ensure alignment to the external standards. However, with the publication of the Manual the logical step for the organisation was to register as a case study for operationalising the concepts and processes encapsulated there.

A secondary aim of the project was to provide feedback to the Council of Europe on their Draft Manual (2003) which was used as a basis for the methodology.

Methodology

As mentioned earlier, the methodology used in the project was based on the procedures recommended by the Council of Europe in their Draft Manual of 2003. However, as the project progressed a number of changes were made to facilitate the operationalisation of the process. The project adapted the four-stage approach suggested in the Draft Manual:

1. Familiarisation
2. Specification
3. Standardisation
4. Validation

In terms of the methodology used, a number of important recommendations were made, these related to the nature of the process (which we suggest is iterative rather than linear as implied in the Draft Manual) and the notion of embedding the process in the institution's test development cycle.

Summary of the Main Findings

The main findings of the project can be summarised as follows:

1. It was found that in order to claim a link to the CEFR at Level B2 the cut score for a passing grade for the Communicator Reading paper should be set at 15 (from a maximum of 30). The same cut score was recommended for the Communicator Listening paper. This is actually in line with current practice for Communicator.

2. Passing levels for the Communicator Writing paper were found to be in line with the Council of Europe recommended tasks for CEFR Level B2. The recommendation is that the cut level for this decision should not be altered at this point in time.

3. The linking process is long and demanding, both at the individual and institutional level. The complexity of the design means that it is expensive for any institution to undertake, certainly to the extent undertaken by City & Guilds in this project. While this perhaps explains the reluctance of many examination boards to undertake a full linking project, we nevertheless recommend that the process be extended to as many of the other examinations in the ESOL suite as feasible.

4. Unless the test which is the focus of the linking project is shown to be robust in terms of quality and level, there is no point in even starting a linking project, as the process is unlikely to succeed beyond the standardisation stage without serious issues emerging. In fact, we feel that with a more demanding specification phase, issues should emerge more clearly at this early stage.

5. Limiting the validation evidence to estimates of internal and external validity is far too simplistic a view of validation. The CEFR should be demonstrated to impact on all aspects of the test, from the test taker to the task to the psychometric qualities and relative meaning or value of the test score.

Based on this project, it is the belief of the project team that the evidence presented here supports the claim that the Communicator tests English ability at CEFR Level B2.

We feel that the process of linking the Communicator examination to the CEFR, has resulted in systematic and sustainable improvements to the test and to the system that supports the test.

It is clear to us that the process has resulted in a test that is more clearly at level, is sound from an internal psychometric perspective and is more replicable and of a high quality. However, that is not all. The systems that support the examination have also been systematically improved and more explicitly linked to the CEFR. The item writers' guidelines are, we believe, up-to-date and more robust than in the past. The specifications are now more likely to result in accurate replication of tests on level – one criticism of the old specification was the lack of detail and exemplification, this appears to have led to a tendency to drift away from the level. This is a warning for other test developers, who take time to specify their tests but do not routinely review these specifications (and their use) to ensure that there is no level or construct drift.

We now feel that we are in a position to consider suggesting a number of Communicator tasks to the Council of Europe for use as recommended level indicators in future linking projects.

Professor Barry O'Sullivan, CLARe
Rachel Roberts, City & Guilds
October 2008

Part 1 – Background to the Project

The City & Guilds International ESOL (English to speakers of Other Languages) examination suite contains two English proficiency examinations set at 6 different levels. International ESOL tests a candidate's reading, writing and listening skills and International Spoken ESOL tests a candidate's speaking skills. Development of the suite started in 2001 and both examinations were launched in 2004.

Given the increasing importance and high profile of the body of work around the CEFR, the decision was made early on in the development process to align the levels of the examination with the levels of the CEFR. The additional clarity offered by the descriptors in the CEFR would add transparency to the assessment system and facilitate stakeholders' ability to interpret the meaning of their learners' results.

The suite of examinations was therefore developed using the CEFR (Council of Europe 2001) as source document to inform the assessment tasks, specifications and assessment criteria. During the development phase the Draft Manual (2003) for relating examinations to the framework was not in existence, so the organisation embarked on a series of internal activities to ensure alignment to the external standards. However, with the publication of the manual the logical step for the organisation was to register as a case study for operationalising the concepts and processes encapsulated there.

It was evident from the start that there would be a significant amount of work and resource required to complete the case study and validate the International ESOL examinations' link to the CEFR. Ensuring that alignment to the levels was ongoing and CEFR methodology was imbedded into our quality process was also a priority. The decision was therefore taken to focus the case study project, and related report, on the alignment process of the most popular level in the examination suite, B2. This report therefore deals with the CEFR mapping project for the International ESOL and Spoken ESOL examinations at Communicator level (B2). The organisation's work to ensure alignment of the other levels has continued simultaneously, but this will not be the focus of this body of work.

In order to complete the International ESOL Communicator level CEFR mapping project City & Guilds has worked in partnership with the Centre for Language Assessment Research (CLARE) based at Roehampton University, London. This has ensured that there is both the necessary expertise to interpret and apply the principles

described in the manual and an impartial perspective on the examination system which adds value to the process.

1.1. The Purpose of the Project

The primary purpose of the project was to gather evidence in support of our claims that the intended level of the Communicator test development team was, in fact, that of the operational version of the test. Therefore, the primary motivation in undertaking this project was to

- Provide evidence that candidates passing the Communicator are likely to have reached Level B2 of the CEFR in order to support claims of the validity of the test.

Clearly this also reflected other expectations of the institution, these can be summarised as:

- That the process of linking, where it included a formal critical evaluation of the test, would contribute to the professionalization of the institution through the development of specialised skills.
- Where the process of establishing a link to the CEFR is embedded in the organisation, future test development projects as well as test validation projects will be facilitated.
- That the commitment of City & Guilds to using the CEFR as the basis for developing all of its English language tests and to establishing evidence of a direct link between its tests and the CEFR through projects such as this will be seen by the wider assessment and education communities as evidence of its commitment to transparency and professionalization.
- That publically available and transparent evidence would greatly enhance the market value of the test, particularly in markets where expectations of empirical support of test-related claims (including level) is more sophisticated.

1.2. The Communicator Test

It was agreed in 2001 to update and redevelop the existing Pitman ESOL and SESOL qualifications with the following objectives:

- To provide assessment at six levels (rather than the existing five) benchmarked to international standards: the Common European Framework
- To develop a suite of awards for the international market
- To improve validity and reliability
- To modernise the assessment while retaining the best features of the existing model

To achieve these objectives a Development team of EFL experts were appointed to:

- Draw up a coherent framework of level descriptors based on the Common European Framework
- Calibrate the existing examinations to the Common European Framework
- Ensure the development of revised test specifications based on the CEF descriptors

The Development team started this project with the aim of providing evidence that candidates passing the City & Guilds Communicator examinations are likely to have reached Level B2 of the CEFR, in order to support claims of the validity of the tests.

A great deal of work involved explicitly linking the B2 Communicator tests to the CEFR Level B2 during a thorough specification process.

1.2.1. Structure of the Test

City & Guilds International ESOL examinations consist of two English proficiency suites, each set at six levels. One suite is focused on speaking (Spoken ESOL) and the other on listening, reading & writing (ESOL).

Development of the suite was completed in 2004. The decision was made early on in the development process, to align the levels of the examination with the levels of the Common European Framework (CEFR), see Appendix for an outline of how each task in the Communicator was designed to meet the expectations of specific CEFR

descriptors. The IESOL examinations are available at six levels benchmarked to the CEFR.

For the IESOL examination, assessment is by a single paper comprising three sections which cover listening, reading and writing.

For the ISESOL exam, assessment is a single exam paper, consisting of a series of questions and situations in which the candidate converses with the interlocutor. Learners can progress through the levels of either suite concurrently or separately.

The listening and reading sections of the IESOL exam paper are marked as an absolute test, in reference to specified answers in a mark scheme with a fixed pass mark for each section.

The writing section of the paper is graded according to a set of assessment criteria covering analytical criteria which are based on the CEFR descriptors for B2. These criteria are task-specific, meaning that an individual scale has been developed for each task in the writing paper (see Appendix 6 for an example of such a scale).

1.2.2. Intended & Actual Audience

The purpose of this qualification is to assess the English proficiency of a non-native speaker of English, focusing specifically on listening, reading, writing and speaking skills. It is aimed at candidates studying English for use in an international environment and for those who want to work/study in the UK, who need externally recognised certification of their level.

The examinations are designed primarily for adults and young adults. For the pre-16 age group, although the examinations are not targeted at this age range, the syllabus, standards and content are carefully controlled to ensure it would not specifically exclude this group of candidates.

Part 2 – The Scope of the Project

The project was initially intended to provide evidence in support of strong claims of a link between the Communicator examination (developed from the beginning with CEFR level B2 as its basis) and Level B2 of the CEFR. With this in mind, the project includes in its design all four stages described in the Draft Manual (2003), these were Familiarisation, Specification, Standardisation and Validation.

The developers of Communicator and sponsors of the project (City & Guilds of London) were also keen that the project would contribute to the professional development of its ESOL staff and also to the overall quality of its examinations. The potential for the process of putting together and operationalising a complete linking process has been neglected by the testing community, though it has always been seen by City & Guilds, and by its partner in the project, the Centre for Language Assessment Research (CLARe) at Roehampton University, London, as representing perhaps the most important long-term value of such an activity.

In fact, the project described in this report is planned to be the first in a series. Since the Communicator is just one examination in a suite of six, the others being Preliminary (A1), Access (A2), Achiever (B1), Expert (C1) and Mastery (C2), City & Guilds plans to apply the lessons learned, and expertise gained in this project to establish empirical evidence of links between all tests in the suits and the CEFR.

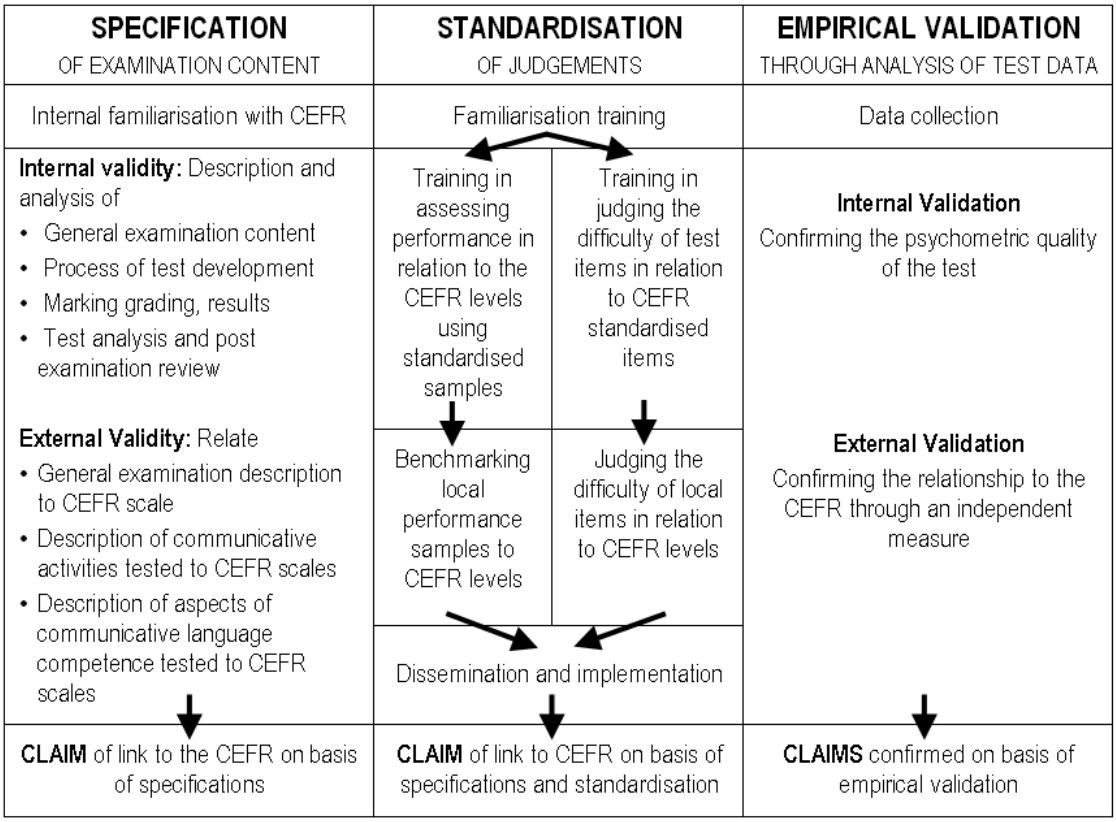
The size and complexity of the project makes reporting its design and outcomes quite problematic – the reports on other projects available at the time of writing focus on a limited linking project, see for example the various reports on the TOEFL and TOEIC reports by Tannenbaum & Wylie (2004, 2005, 2007), the iBT project, described in brief in the Executive Summary (ETS, 2007) and the Trinity College London report (2007) and are thus relatively straightforward in the way they present reports.

Since this project marks an early attempt to apply all four of the recommended stages in the Draft Manual (as outlined in Figure 2.1), the situation is different. For this reason, we will present each stage as an individual element of the study – with a description of the method (including description and discussion of the participants, the instruments) together with the results and related claims. Finally, we will present an overview of the project as a whole, focusing on the iterative nature of the whole process, the implications

of the work done, its limitations, as well as presenting some recommendations for the manual developers and users.

The lessons learnt during the process have made a significant contribution to City & Guilds as an institution. The final chapter will begin to explore this contribution and to suggest that the process of linking can serve a number of purposes to an examining board.

Figure 2.1. The Original Linking Model (Draft Manual, 2003: 4)



Part 3 – The Familiarisation Stage

According to the Manual the objective of the Familiarisation Phase is to ensure that the participants in the process are aware of the details and interpretation of the descriptors "[B]efore embarking upon the description of the examination." (Manual, 2003: 6)

As previously stated, the City & Guilds International ESOL examination suite was designed to align the levels of the CEFR. Therefore, certain familiarisation activities were initially carried out during the development phase of the examination. The objectives of this phase of the case study project has been to:

- briefly review the familiarisation work that took place during the development phase
- plan and document the ongoing work that has been informed by the recommendations in the Manual, that is taking place to ensure that all key stakeholders have an in depth understanding of the frameworks.

3.1. Development phase

In order to capture the nature of the familiarisation work done during the development phase of the examination i.e. pre case study project, a questionnaire was developed to canvass the views of the consultants and examiners who worked on the development of the exam. They were asked to list the nature and value of the familiarisation activities they completed. From the evidence collected, we know that the activities included: sorting the CEFR descriptors, self assessment of language levels and benchmarking local samples of candidate speaking and writing performance to the CEFR levels. However the results of the questionnaire and anecdotal feedback illustrated that even with this degree of familiarisation, there was still some doubt about the true meaning of the levels and how they actually translated into language ability of a "real" learner.

3.2. Ongoing familiarisation

The experience during the development phase demonstrated that in order to ensure the model of language competence represented by the CEFR is internalised appropriately, work on CEFR familiarisation need to be in depth and ongoing. Therefore work has been done to ensure that the CEFR model of language has been embedded into staff

development and consultant training. Without this degree of familiarisation it becomes exceptionally difficult to make reliable decisions on levels during the specification, benchmarking and standard setting phases. Given that many of the assessment professionals working for our examination also come into contact with a range of other assessment criteria and standards, it seems obvious that unless the CEFR has become second nature, there will be a degree of interference from the other standards in question.

It is also apparent that all stakeholders working on the project would need an in-depth understanding of the examination system. Easy for the consultants and staff who work with the examination on a day to day basis, but essential too for the external consultants who were drafted in to maintain a level of external scrutiny and impartial perspective.

According to the guidelines in the Manual CEFR familiarisation activities have been conducted with all the key stakeholders involved, including:

- examiners for both writing and speaking exams
- examination development consultants
- item writers
- external consultants - involved in the benchmarking and standard setting phase

This involved the recommended activities described in the Manual such as sorting of descriptors, self-assessment of language level etc. However, there have also been a number of other steps taken to increase the effectiveness of the familiarisation training and adapt it to the needs of the organisation. This involved:

- The key stakeholders have been supplied with copy of the 'blue book' (Council of Europe, 2001) in order to become familiar with the complete model and not just a small cross section of the descriptors.
- Familiarisation training has been carried out with the City & Guilds staff responsible for managing and administering the assessments and project. The rationale being that the methodology and meta-language associated with the CEFR permeates throughout the organisation and constantly being discussed, reiterated and embedded.
- The training typically involves a self-study preparatory phase that can be completed pre-event. Instead of time-tabling in what is thought to be appropriate time for the

activities, they can be completed at each individual's pace. The activities used are based on those suggested in the Manual (consideration of questions contained in the relevant Chapters of the CEFR; self-assessment of own level of at least one language other than English).

- The activities used in the actual training event also reflect those mentioned in the Manual (discussion of CEFR levels as a whole using the global scale; sorting and reconstructing the descriptors from each CEFR scale)
- Familiarisation Training is a precursor to the item writer training sessions.
- Time has been spent building the expertise of a fixed core group of consultants that the organisation has sought to keep stable. The rationale being that there is likely to be agreement and mutual understanding on the true nature of the levels. The consultants are also going to have a true picture of the exam. This is unlikely if they are only drafted in for a fixed period on an occasional basis.
- Familiarisation is a fixed feature of item writer recruitment and development.
- Familiarisation activities also form part of the regular examiner training sessions, which take place at least one per year

As well as the formal training, additional tools have also been devised to continue the familiarisation process outside the training room and ensure that it is iterative. These include:

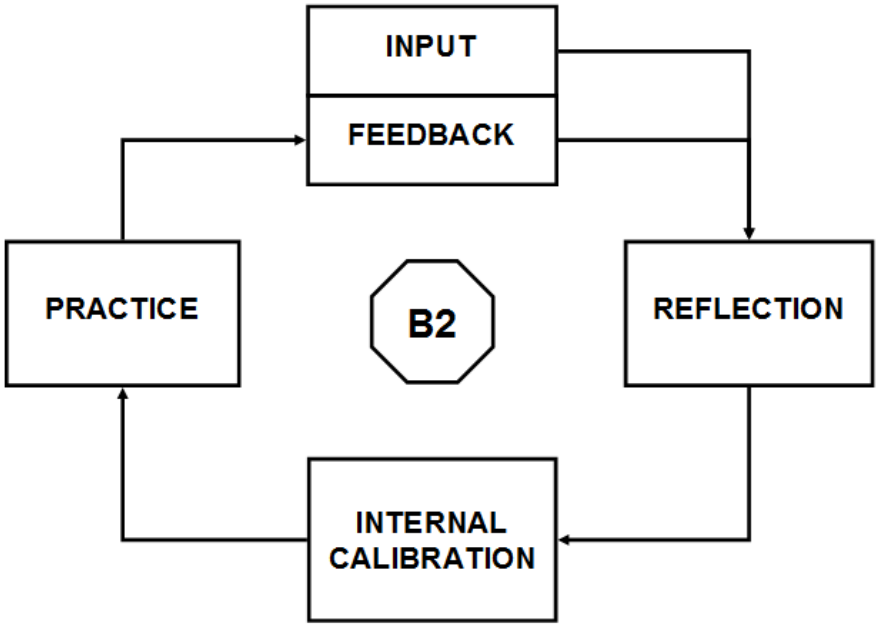
- documentation support – item writer guides, Blueprint for exam and marking guides includes specific references to the descriptors. Which make the links very explicit and continually reinforce the standards
- syllabus – grammar lists, function lists to add help to further define the levels, take away the fuzziness
- through research into text features additional insights have been provided to item writers into the features of input texts at certain levels
- the use of external CEFR experts to give impartial feedback to the organisation on the interpretation of the CEFR standards
- feedback loops – both qualitative and quantitative

The work needs to be ongoing to maintain a stakeholder involved in developing the exam or validating the level individual on the path towards truly understanding the levels.

3.3. Iterative familiarisation

With training, feedback, reflections and practice, ideas of the standard converge towards a common understanding of a particular level. Without the iterative nature and ongoing familiarisation there is a tendency for divergence i.e. a moving away from the common understanding of the level. For this reason we developed a model of familiarisation for all City & Guilds test development and linking projects. This model can be seen in Figure 3.1.

Figure 3.1. City & Guilds Model of Familiarisation



The model of familiarisation outlined above incorporates three steps

- Communication
- Documentation
- Systematisation

Communication describes the objective of making sure that everyone involved in the process is speaking the same language. This is achieved by involving everyone in the training helped to embed, reiterate, ensure the language was the norm and constantly used reinforcing both the level and, more importantly, the interpretation of the level for the examination in question. In practice, it is reflected in such ways as extra support offered to item writers to identify key issues in creating appropriate tasks and items for a specific level, or in the systematic research-driven evaluation of reading tests which is used to identify suitable texts at each CEFR level.

Documentation The CEFR reference were incorporated into the materials, so the people who are engaged in all aspects of the development and operationalisation of the examinations are constantly referred back to CEFR levels

Systematisation feedback loops – working to make sure that familiarisation improves over time.

Within the model, we see that the input (in the form of the various pre and during-event tasks referred to above) leads to a period of reflection. During this stage of the familiarisation process, the participants are encouraged to re-visit the documentation and to consider how the lessons learnt from the input stage have affected their understanding of the level in question, in the model shown, we have placed the level which is the focus of this project at the centre, obviously, this would change depending on the level of the examination being linked. We believe that it is only following this period of reflection that the participants attain internal calibration. In other words, they have internalised the level. At this stage, participants are encouraged to apply their understanding of the level to the practice of test task development and/or standardisation. The lessons learnt from the success or failure of this application of the knowledge built up during the earlier stages then feeds back directly into the entire process, as the input materials and approach are routinely evaluated and updated where necessary.

Part 4 – The Specification Stage

In this part of the report, we will outline the procedures, instruments and outcomes of the Specification stage. The section will end with an outline of the claims we feel we can reasonably make regarding the validity of the Communicator examination based on the evidence presented here.

We should make it clear from the beginning that we do not feel that even weak claims of validity can be made from the type of evidence provided at this stage of a linking project. This is because the quality of the evidence required to make these claims is, in our opinion, quite weak in itself as it is generated by self-assessment with no corroboration from other (outside) sources. We feel that claims made during the process are important, but should only be used as evidence that a stage has been successfully completed and that the linking process can proceed to the next stage.

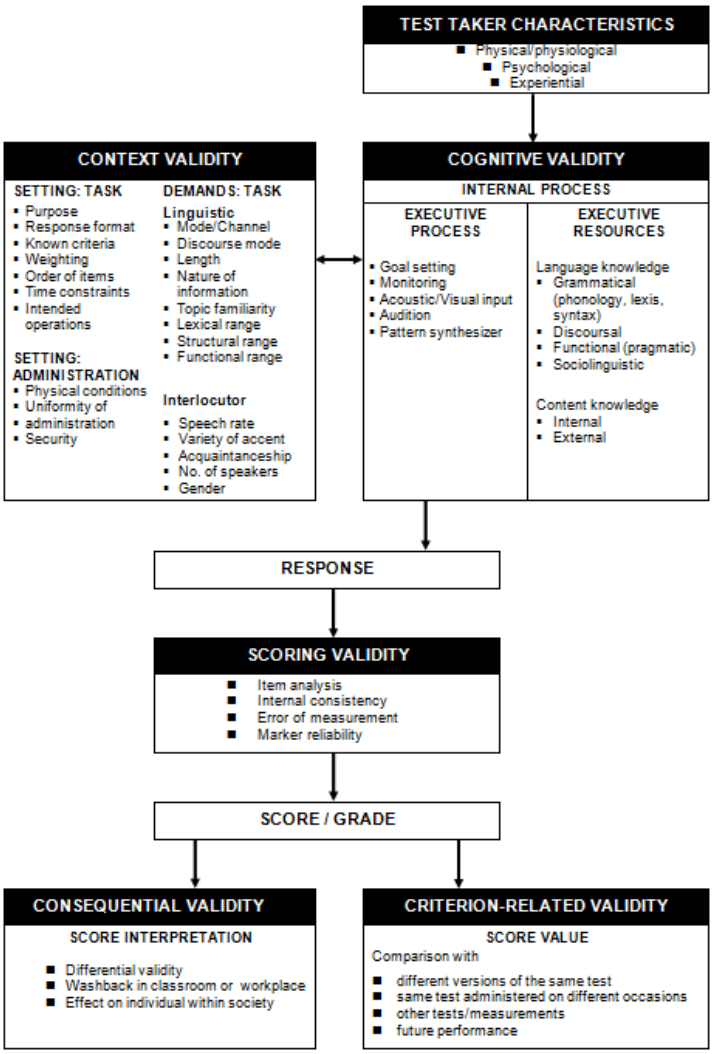
4.1. Preliminary Decisions

The original specifications, which were based on a detailed Test Syllabus (summarised in Appendix 1) that had been developed to link directly to the CEFR descriptors of the three skill areas (Listening, Reading & Writing), were found not to be working as expected. By this we mean the specifications did not contain the level of detail required for replicability of test versions at the CEFR level indicated in the syllabus. For this reason, the decision was taken to return to the specifications with a view to re-writing them in a way that made them more user friendly and more likely to result in consistent test papers.

To do this we used the validation frameworks which were developed at Roehampton University, London, published by Weir (2005) and operationalised for use as a basis for test specifications by O'Sullivan in a number of test development projects (e.g. QALSPELL, 2004; Exaver, 2005). These frameworks offer a broad description of the test from the perspectives of the **test taker** (includes an explicit description of the test population from the perspectives of physical, psychological and experiential characteristics and a model of the cognitive processes and resources required of the test taker), the **test task** (referred to by Weir as *Context Validity* – involving a systematic description of the test task taking into account a range of parameters and from the perspectives of performance conditions, language operations and administrative setting)

and finally the **scoring system** (or *Scoring Validity* according to Weir – a systematic description of the procedures to be followed in order to arrive at a valid score).

Figure 4.1. Framework for Listening Test Validity (from Weir, 2005)



In the original work by Weir, there are four separate frameworks (one for each of the skill areas). An example of how the context validity element of the listening framework (Figure 4.1.) was used to re-specify a task in Communicator listening paper can be seen in Figure 4.2., below. When the test had been fully re-specified, notes were taken of tasks and items that were potentially problematic (in that the original specifications were somewhat unclear and where there was some possibility that the resulting tasks and/or items may drift off level). With this part of the process now completed it was decided to proceed to the completion of the specification forms. This process is described in the following sections.

Figure 4.2. Example of Task Specification based on Weir's Validation Framework

TASK TYPE	**Description**		
Format	MCQ Three short, unrelated dialogues.		
Purpose	Listening to identify specific aspects of spoken dialogue 1. Purpose 2. Context 3. Gist 4. Attitude 5. Relationship between speakers 6. Speaker's feelings & opinions 7. Function of utterance		
Response Format	MCQ – four options		
Known Criteria	NA		
Weighting	All items are weighted equally		
Time Constraints	30 seconds allowed for response (to start of next item)		
TEST DEMANDS			
INPUT			
No. Inputs	3 dialogues		
Input Focus (see pages 53-57)	Dialogue 1 Topics 1, 2, 4, 5 or 9	Dialogue 2 Topics 3, 6, 7, 8 or 12	Dialogue 3 Topics 10, 11, 13, 14 or 15
	Note: all dialogues should reflect authentic language use **and** include at least one piece of distracting information		
No. of items	6 in total – 2 items per input text		
Item Focus	Dialogue 1 Any two from 1-4 (see *Purpose*)	Dialogue 2 Any two from 1, 5 or 6(see *Purpose*)	Dialogue 3 Any two from 2, 3, 5 or 7 (see *Purpose*)
	NOTE: try to avoid including a focus in more than one dialogue		
Channel	Aural only – Heard **twice**		
Discourse Mode	All interactional		
Text length	5 to 8 turn exchange, maximum of 110 to 160 words per dialogue		
Item Length	Stem – maximum of 12 words Options – all approx. same length – max 30 words total (e.g. 4 x 7 words for each option)		
Nature of information	Concrete		
Structural Range	Choose from GRAMMAR LIST (B2)		
Functional Range	For Dialogue 3 select function from FUNCTIONS LIST – all options must be from this list		
CEFR B2 descriptors	***B2 General Descriptors*** • *Can understand the main ideas of propositionally and linguistically complex speech on both concrete and abstract topics delivered in a standard dialect.* • *Can follow extended speech and complex lines of argument provided the topic is reasonably familiar, and the direction of the talk is sign-posted by explicit markers.* ***Understanding conversations between NS*** • *Can keep up with an animated conversation between native speakers.* • *Can with some effort catch much of what is said around him/her, but may find it difficult to follow effectively a discussion with native speakers who do not modify their language in any way.* ***Listening to Audio media and recordings*** • *Can understand recordings in standard dialect and identify speaker viewpoints and attitudes as well as the information content.* • *Can understand most recorded audio material delivered in standard dialect and can identify the speaker's mood, tone etc.*		
SPEAKER			
Speech rate	Average native speaker speed (approx. 120 to 140 wpm)		
Variety of accent	Standard native speaker accent		
Relationship between speakers	Vary – equal and unequal (item writer's see Relationship List) Should be approximately equally represented where relationship is tested		
No. of speakers	2		
Gender / Profile	Male and female		
Expected Output			
Channel	MCQ		

4.2. Forms completed

The set of specification forms presented in the Draft Manual (Council of Europe, 2003) were completed for all three papers. A copy of the completed forms can be found in Appendix 2.

4.3. Procedure

The procedures followed in completing the specification forms are briefly outlined in this section. The process adapted when completing the forms was, like much of the work on this project, iterative in nature.

The forms were initially completed by a team of three people from the City & Guilds ESOL group. Each member was first asked to complete sections that related directly to their area of expertise, for example the productive tables for writing were completed by the person whose responsibility it was to create these papers. The chief examiner was also asked to be part of this group, offering her experience in writing the original specifications for the Communicator. When the individual elements had been completed, the team met as a group to try to reach consensus on the completed forms. Where consensus could not be reached, the varying positions were recorded on the form and highlighted.

At this point the forms were sent to CLARe for additional input. This input consisted of the CLARe team completing the forms based on our understanding of the Communicator (from the handbook, specifications and sample test papers supplied by City & Guilds). These completed were then compared with the originals and any differences noted. Finally, a meeting was held at which the differences were discussed by the team members from City & Guilds and CLARe. The wording for all tables was agreed on at this point.

Because the procedures used in this project called for an additional stage, a preliminary critical review of the test, the specification stage was re-visited after the updates to the Communicator had been agreed, trialled and implemented.

The procedure at this point was the same as for the original form completion stage described above. An internal team first re-visited the forms taking any changes into consideration. As before, this was done in two stages, with each member initially working

alone and later getting together to reach a consensus. Meanwhile the CLARe team repeated the process before finally the two teams got together to complete the forms as they are to be found in Appendix 2 of this report.

4.4. Lessons Learnt

It was felt that a number of invaluable lessons were learnt during the specification stage. Among these lessons we felt that there was a deepening the knowledge of the CEFR levels among the participants, a greater awareness of the need to constantly look to the CEFR rationale and descriptors when developing and writing items and a broadening of the institution's understanding of the concept of quality. However, it may be most valuable to the readers of this report if we focus at this point on what we felt were the two main lessons learnt, these relate to the quality of the working specifications and to the use of external expertise at this stage of the project.

As with the other phases of the project, the specification phase was initially expected to be essentially linear in structure and relatively straightforward in practice. The most obvious lessons learnt from the actual procedures was that the formal specifications of the test in question were not expressed in as clear a manner as they might have been while the completion of the specification forms (A1 – A21) was both time-consuming and at times unnecessarily complex. The positive side of this complexity was that the original specification were scrutinised in a way that had not occurred before this project. It was decided by City & guilds that the deficiencies in the specifications had to be addressed before the project could continue. In the process of doing this, it became clear that there may be some areas within the Communicator that might prove problematic and that a critical review of the entire test should be carried out. As a result, this additional phase was added to the project. The lesson here is that unless the test developer is convinced that the test is at the appropriate level and of a sufficiently high quality, any linking project is either bound to fail or to result in meaningless claims. Of course all developers will argue that their tests are working perfectly well, though without some level of unbiased critical appraisal this may not actually be the case.

One of the advantages to using an external source such as CLARe, was that there was less likelihood that judgements would be affected by preconceptions of test level or quality. Another advantage was the experience of the CLARe team with other linking projects – one in Turkey (Bilkent University COPE linking project) and a second in

Mexico (Veracruz University/Roehampton University EXAVER linking project). Because the three projects (Communicator, COPE and EXAVER) were at different points in the process, experience gained from one fed into the other projects. We were, for example, able to learn from the teams in Ankara and Veracruz who were working on the forms about how they interpreted certain items and how and why the answered items in particular ways. This sharing of ideas and materials proved invaluable to this project, though we recognise that many examination boards would be very reluctant to adopt such a strategy.

4.4. Some Observations on this Process

While at first the forms seemed somewhat awkward and repetitious, the process of completing them was useful as it forced the team to consider aspects of the tests not necessarily referred to directly in the re-written specifications.

An example of this was the notion of interactive and productive writing. In the specifications no difference was seen between the two as the writing tasks were essentially seen as being productive in nature. Traditionally, it is assumed in such tasks that by ensuring that each writing task has a clearly described audience the candidates' awareness of this audience would encourage them to take into account the interactive nature of the text they were about to produce. The tables relating to interaction prompted some debate about the interactive nature of the writing tasks reviewed and led to the realisation that the fact that the rating scale used for the tasks actually include an element of interactivity, made it important that the individual candidates are made fully aware of the need to take the eventual reader of the work into consideration at all phases of the writing event. This realisation led to changes in the rubric to ensure that candidates were made aware of the

However, we feel that the design of the manual forms should be reconsidered to reflect an up-to-date understanding of test specification and validation. We used the validation frameworks suggested by Weir (2005) and feel that these offer one obvious solution as they are very practical as well as being theoretically sound.

4.5. Evidence and Claims from this Process

The procedure adopted to ensure that the manual forms were completed in an accurate and systematic way (through a series of critical discussions and reviews employing both internal and external experts) resulted in a set of completed forms that, in the opinion of the City & Guilds and CLARe teams, accurately reflect the Communicator papers.

As can be seen in the completed forms (Appendix 2), the Communicator appears to reflect the original expectations of the developers, in that the conclusions from the various forms indicate that the level of the examination is CEFR B2. The evidence from the form completion exercise suggests that the test papers are a true reflection of the CEFR B2 descriptors for the three skills tested.

The fact that there are clear working specifications that have been devised to reflect the original design and level of the test that are the basis for task and item writer training and monitoring is evidence that City & Guilds as an institution is committed to ensuring that the CEFR is embedded into the test development cycle and that the quality and level of the Communicator reflects this commitment.

While the Draft Manual (Council of Europe, 2003) suggests that it should be possible at this point to make claims based only on the test specification, we feel that any such claim at this point is likely to be premature and possibly even meaningless.

The evidence gathered so far is based almost entirely on the developing institution's vision of the test papers. We could at this point take quotations from the completed forms as evidence, for example, that the Communicator listening paper addresses specific elements of the CEFR B2 descriptors for listening. However, even here the interpretation of this evidence is essentially institutional – though we did try to counteract this by including external experts in the process – and cannot really be seen as concrete proof that the paper really does reflect the level it is claimed to reflect. Since the vision of the institution will be in some ways biased, or at least it will not be impartial, we feel that only a very weak claim of test level should be made based on the evidence contained in the completed forms. Instead, we would argue that the claim we make at this point is sufficiently strong to allow us to progress to the next phase of the project, in which evidence of a more empirical nature will be presented in order to demonstrate that the critical boundaries for all three papers reflect the expectations of users of the CEFR.

The section of the report that follows, therefore, outlines the procedures and results of a set of standard setting events set up by City & Guilds to gather information that will

allow the institution to make claims that the critical (i.e. pass/fail) boundaries of the papers are linked to CEFR Level B2.

Part 5 – The Standardisation Stage

During this stage, it was decided to use, from the beginning, a group of expert judges who represented both the institution and the outside world of language learning and assessment. This decision was made as it was felt that a group comprised solely of insiders might suffer from any one (or more) of a number of biases.

In the first of these cases, there is a real possibility that those who were involved with the original development might base their decisions on the *intended* level of the test rather than on the *actual operational* level. This is particularly true of a test which has been written specifically to a given CEFR level, and Communicator is one such test.

In many ways, this stage can be seen as the core of the project. At this point, evidence is gathered that should (all going well) allow us to make stronger claims of a link to the CEFR than we were in a position to make following the Specification Stage. Here, we examine the operational examination (by using live test tasks) as opposed to the specifications. This is important as we can never be certain of the way in which the item and task writers interpret the specifications until we first see the resultant items or tasks and gather evidence (both qualitative and quantitative) of their appropriacy in relation to the intended level.

One of the important aspects of this stage of the project was the decision to look at the test in terms of its likely link to the CEFR independently of the actual standardisation. This was done as it was felt that there would be no real point in setting pass/fail boundaries in relation to the CEFR (the objective of standardisation) if there was evidence that the level of the test, or part of the test might be problematic. This is, in fact, in keeping with the argument presented in the Draft Manual (2003: 66) that the test must be valid and reliable before any meaningful link can be claimed. Since any validity argument will include evidence that the items or tasks are actually testing candidates at the correct level, we felt that it was first necessary to establish that the tasks were 'on level' before we could proceed to any standardisation. This decision proved appropriate in light of the findings of the preliminary expert panel (see section 5.2.4 below).

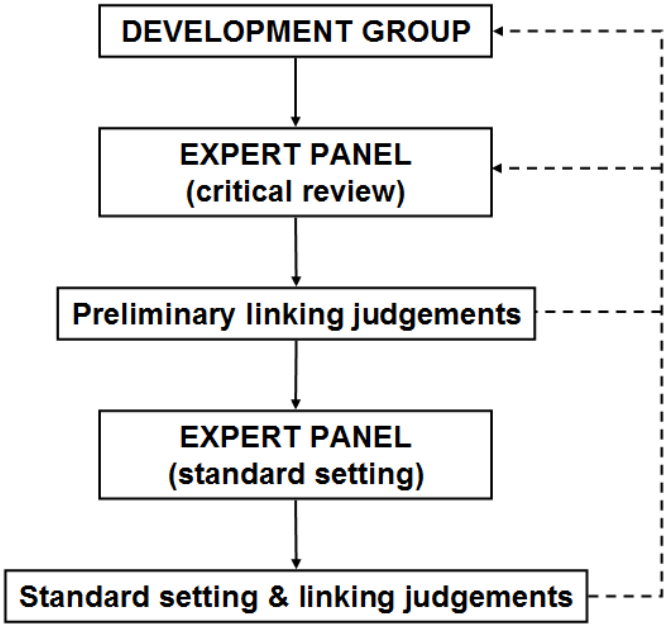
5.1. The Methodology

In this section we present the various aspects of the methodology used to gather evidence in support of the claims we later hoped to make regarding the link to the CEFR. We begin this section by outlining the approach taken.

In order to gain a meaningful insight into the standardisation process, which had as its aim the setting of pass/fail boundaries which reflected the border between the minimally capable B2 candidate and their less than capable peer, we first planned each of the standard setting events to include pre-event work as well as the work to be undertaken during the event. In order to achieve this, the following design was agreed and implemented.

The test development group within City & Guilds, who are responsible ultimately for all aspects of the development and validation process, worked with the project partners (CLARe) to decide on the details of the process. These included the setting up of an expert panel to make preliminary linking judgements, the selection of test tasks which were seen as representative of the Communicator level, the preparation of the preliminary work to be carried out by the experts prior to meeting and of the procedures for the meeting. Finally, the group was expected to deal with any suggestions made by the expert panel in its preliminary meeting. In this meeting, the panel was to review the current test tasks in light of the CEFR and also with regard to overall quality. Decisions made by the panel might be that all was in order and that the next stage (i.e. standard setting) should be conducted or that there were aspects of the test that should be seen to be appropriate in terms of quality and approximate level before such a step could be recommended. In its final embodiment, the expert panel would be convened to first review any changes (where they have been recommended) and then to set the pass/fail boundaries in relation to the CEFR B2 level. Figure 5.2. shows the iterative nature of the planned process.

Figure 5.1. The Design of the Standardisation Process



5.2. The Expert Panel (Critical Review)

We decided to employ an expert panel of six people to look at the tasks in relation to the CEFR. While it may appear that this is a relatively small group, it should be understood that no standardisation decisions would be made by this group. Instead, it was expected that the group would use their expertise to make judgements on the level of the various tasks presented to them in relation to the descriptions of Level B2 from the CEFR documentation.

5.2.1. Objectives

The primary objective of the standard setting activity was to set the pass/fail cut score for the Communicator examination that could be shown to correspond to a commonly understood interpretation of the border between a minimally capable B2 level candidate and a candidate who narrowly falls short of that standard.

Since the project team were aware that the Communicator examination itself would first need to undergo a stringent exploration of its qualities, a second (though not secondary) objective was identified. This was that the test tasks could be demonstrated as being at the appropriate level and also of sufficiently high quality. It was this second objective that was focused on by the first expert panel.

5.2.2. Composition

As indicated above, the expert panel was actually not a single body, but in fact was two separate groups. The first of these panels was expected to look at the examination in terms of CEFR level and overall quality, and to make recommendations regarding the next stage of the project. In other words this panel was expected to undertake a critical review of the Communicator. This part of the process is discussed in the current section. The second, larger, group would later make final linking and standard setting decisions based on an examination that had already been through a systematic qualitative review. The activities of this group are reported in Section 5.2.3. below.

The expert panel for the critical review was comprised of six members. Three of these were insiders, one of whom was directly involved in the development of the original specifications, and continued to work with the item writers who interpreted these specifications to develop new items and tasks. All three of the insider group had extensive experience in working with the CEFR for assessment purposes at the B2 level. The remaining three members were unconnected to City & Guilds in any way. All three were experienced language teachers at this level, and two had significant experience in test development again at the level. All of the panel members had experience of standard setting, though the range of this experience varied from some participation in such an event to organizing these events on a regular basis.

The reason we felt it prudent to opt for such a balance in the group was our concern that a panel comprised only of insiders might display some bias in their judgements due to their involvement with the test and/or the company. It was also recognized that this bias might be positive or negative, depending on an individual's involvement with the institution or with the test development project itself. It was felt that no panel could hope to perform the work we were asking without a core of individuals with a high level of what we would term 'local expertise', by this we mean people who were expert in all aspects of the examination (e.g. development, format, candidature, use etc.). The outsider group was put in place to ensure a systematic and unbiased exploration of the examination from a neutral perspective. These panel members were paid for their input.

In addition to the above, the preliminary panel had an independent chair. The role of this person was to oversee the whole process and to make strategic recommendations where appropriate. The role was taken on by a member of CLARe, as it was felt that this person

would have sufficient familiarity with both the CEFR and the examination while not having a strong commitment to the examination itself.

It should also be noted at this point that the panel was asked at its first session to review all of the skill areas (i.e. all of the papers in the Communicator examination). This was done so as to allow the project team to overview the entire test before continuing to the later critical stages of the project.

5.2.3. Procedure

There were six participants in this stage. Of the five, two were members of the City & Guilds ESOL team (one being the Chief Examiner), while the other three were all from outside of the organisation as was the independent chair. All were invited to participate in the preliminary expert panel because of their extensive knowledge of the CEFR (demonstrated through their previous participation in teaching and/or examining at level B2. The participants were sent a package of information and asked to

a) re-familiarise themselves with CEFR levels B1, B2 and C1 (all participants were chosen based on their demonstrated familiarity with the CEFR through their work on other projects – this re-familiarisation was considered by the project team to be a useful exercise as it was expected that such an activity would have a similar effect on the participants as was found by Rethinasamy (2006) who demonstrated that even a brief review of test materials can have a significant and positive effect of rater behaviour),

b) familiarize themselves with a series of tasks from Communicator Listening, Speaking Reading & Writing papers (in the case of the insider group this would have also been a re-familiarisation exercise),

c) review the performances (writing and speaking, though the latter came from a separate stand-alone examination aimed at CEFR Level B2 and designed to augment the Communicator examination) and exemplar tasks provided by the Council of Europe as part of their standardization CD-ROM,

d) estimate the CEFR level of a set of responses to each of three performances on four Communicator writing and speaking tasks taken from actual test data – using the Council of Europe standardized task performances as a base,

e) estimate the CEFR level of a set of live Communicator reading and listening tasks (10 tasks for reading and 10 for listening).

Almost all of the participants performed these tasks before attending the panel meeting, one of the internal group was unable to perform the final two stages due to pressures of work. Their estimates were submitted to CLARe for analysis. Multi-faceted Rasch analysis was used to gain an understanding of the process and of the ability of the participants to perform the tasks expected of them. This analysis was also used to identify tasks with which judges experienced difficulties. These tasks were then prioritized for the discussion phase of the panel meeting. The program used to perform this analysis was Facets.

Note that we report below the outcomes of a speaking paper review. This paper is not part of the Communicator test, but instead is a test that is independent but also aimed at CEFR Level B2 and is meant to complement the Communicator. The test is also currently being linked to the CEFR.

5.2.4. Outcomes from the Preliminary Expert Panel

The results of the analyses described in the previous section indicated that all of the judges appeared to be consistent in their judgements, though this was most apparent in the case of the productive skills. Figure 5.2. shows a simplified summary of the judging performances for these two areas (for a more detailed Facets output for all these papers please see Appendix 3).

In the following figures, 'level' indicates the measure in logits (in other words it is an estimate of the difficulty level of the tasks in listening and reading or the proficiency level of the performances for writing and speaking). These estimates were found by analysing the responses of the participants using Multi-Faceted Rasch (MFR) analysis. In both cases we can see that the spread is quite small at 1.08 logits (the unit of measurement used in MFR). This suggests that the judges tended to agree with each other when it came to identifying the level of the performances for writing and speaking. While some judges were harsh and others lenient (as we would expect), there are some interesting outcomes here.

Figure 5.2. Summary of the Preliminary Expert Panel Judgements for Speaking and Writing

Judge	Trend	Level	Internal Consistency
OC	Lenient	-.47	Good
OS	Harsh	.61	Good
OD	Harsh	.28	Good
IR	Lenient	-.04	Good
IS	Lenient	-.38	Good

SPEAKING

Judge	Trend	Level	Internal Consistency
OC	Lenient	-.17	Good
OS	Harsh	.11	Good
OD	Harsh	.26	Good
IR	Lenient	-.03	Good
IS	Lenient	-.17	Good

WRITING

SMALL SPREAD	VERY CONSISTENT

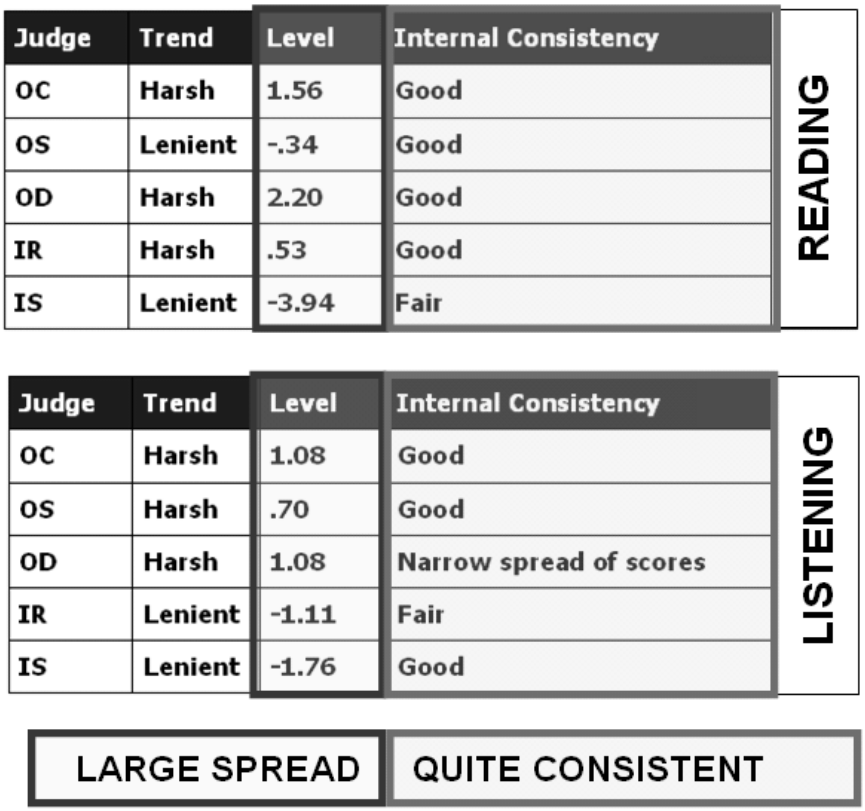
Note: The Judge code O represents an outsider, while the code I represents an insider

One of the interesting features is the confirmation that the internal judges tended to be somewhat lenient in their decisions. In other words, they were likely to view a performance as being on level. However, it should be noted that one of the outsider group (OC) was even more lenient for the speaking and very similar to the insider group for the writing. It should be pointed out that there was a degree of bunching by the judges and a high level of agreement. This was clear from the low spread of the 'level' measure and by the fact that in both cases the fixed (all same) chi-square in the raters measurement report rejected the null hypothesis – suggesting that these judges were delivering similar decisions to a statistically significant extent (see the raters measurement reports for speaking and writing in Appendix 3). It is also interesting to note the high levels of internal consistency of all five judges – see the Infit Mean Square column of the same tables as referred to in the previous sentence.

The results for both reading and listening were quite different. Here (see Figure 5.3.), we see that there is a far greater spread of estimate by the judges (6.14 logits), whose internal consistency was not as high overall as we saw for the productive skills.

Feedback from the judges suggests that they felt at the time more comfortable making decisions based on candidate performances rather than trying to get inside the head of the candidate to try to predict how they might perform on the receptive tasks. This is clearly an issue to be dealt with by the organizers of any standard setting event for tests of the receptive skills.

Figure 5.3. Summary of the Preliminary Expert Panel Judgements for Reading and Listening



Judge	Trend	Level	Internal Consistency
OC	Harsh	1.56	Good
OS	Lenient	-.34	Good
OD	Harsh	2.20	Good
IR	Harsh	.53	Good
IS	Lenient	-3.94	Fair

READING

Judge	Trend	Level	Internal Consistency
OC	Harsh	1.08	Good
OS	Harsh	.70	Good
OD	Harsh	1.08	Narrow spread of scores
IR	Lenient	-1.11	Fair
IS	Lenient	-1.76	Good

LISTENING

This group felt that the test tasks they reviewed for all four skills were likely to result in an overall level B2 performance. However, it was also felt that there were some changes required to ensure that this level of performance could be more systematically achieved.

With the reading and listening it was felt that one task in each paper was operating right at the lower end of the level. For this reason, it was suggested that a slightly more demanding version of each task should be designed and trialled, so that it would be more likely that the overall performance on each paper would more solidly represent the B2 level.

For the productive skills, it was agreed by the panel that the tasks were likely to result in B2 level performances, and that the examples for both writing and speaking clearly

indicated this. However, it was felt that grades awarded to task performances at the main critical boundary (i.e. at the pass/fail boundary) were not always consistent with the working definition of a minimally competent B2 candidate. The Preliminary Expert Panel reviewed the rating scale and found that it was unproblematic, therefore, they decided to recommend that City & Guilds review their rater training procedures to emphasise the critical boundary. It was also suggested that the institution use a number of Council of Europe standardised writing and speaking tasks in its rater training process. This was agreed by City & Guilds and new guidelines for trainers were commissioned. It was also mooted at this point that City & Guilds might move towards using task-specific rating scales for the writing tasks as these would make the process more valid (in that different tasks often have specific expectations with regards to successful performance), would possible result in more reliable (systematic and consistent) marking, and would actually speed up the marking process. This advice was acted on later in the project and such scales devised, see Appendix 6 for an example.

An external consultant, was then asked to work with the developers to look over these tasks and recommendations with the brief to re-specify where necessary and then trial all new task versions.

A report on the results of this work follows.

5.2.5. Task Re-Specification & Trials

Since the productive papers (writing and speaking) were found to be essentially working well, the only recommended changes being to the interpretation of the pass/fail boundary, it was decided that this trial should focus primarily on the receptive papers. The re-specification of the test led to some changes to both the reading and listening test papers. These recommendations were then reviewed by the Preliminary Expert Panel who contributed to the finalization of the tasks. This activity emphasises the iterative nature of the entire linking process, as this type of constant review and evaluation was found at all stages of the project.

While many of these changes were minor, there were some new tasks included in the trial. In addition, it was also decided to ask all trial participants to perform at least one each of the Cambridge ESOL and Finnish reading and listening tasks claimed by the Council of Europe to have been standardized to level B2. This would allow us to gain

some understanding of the psychometric qualities of the tasks and of their level in comparison to the so-called 'standardized' tasks.

In addition to the papers all participants were asked to complete a short self-assessment based on 'Can Do' statements. These statements covered levels B1, B2 and C1 of the CEFR and were to be answered using a five point Likert scale (see Appendix 4). The classroom teachers were also asked to indicate what level they felt each student was currently at (in terms of the CEFR).

The list of tasks trialled at this point is shown in Table 5.2.

The participants in the trial were 59 language students attending a UK university pre-sessional English programme. The participants came from a variety of backgrounds and disciplines and were expected (by their teachers) to perform at a range of level from A2 to C1, though primarily at level B2. One participant failed to respond to many of the items on the reading paper and none on the listening paper so was eliminated from the data, leaving a final population of 58.

Table 5.2. Trial Papers

Paper	Changes
Listening	Task 1 – New task Task 2 – Updated task Task 3 – Updated task Task 4 – No changes Task 5 – Cambridge ESOL Task Task 6 – Finnish Task
Reading	Task 1 – No changes Task 2 – New task Task 3 – New task Task 4 – Updated task (graphics removed from original) Task 5 – No changes Task 6 – Finnish Task Task 7 – Cambridge ESOL Task

5.2.6. Analysis of the Trial Data (Listening)

Though the trial population was relatively small, the number was still expected to give us an indication of the level and performance of the tasks.

In the first table (Table 5.3.) we can see the mean facility and point biserial values for all tasks in listening paper. We can see clearly that there appears to be one easy task (number 4) but that, in general all tasks appear to be working well in terms of mean facility value and point biserial (there is no definite range of acceptability for the facility value, though we would not expect that the maximum range would extend beyond approximately 0.30; on the other hand even if we select a conservative expectation of anything over a point biserial of 0,30 as being acceptable, then all tasks meet the requirement). In other words, all tasks are relatively similar in terms of difficulty and are discriminating well between the more and less able candidates. However, it should be noted that Task 4 appears to be out of the range of the other five tasks, in that it is quite a lot easier. .

In addition to the above, scale analyses was performed in SPSS. As part of this process, it is possible to request an estimate of Cronbach's *alpha* for the test if each item in turn is removed. If the estimate goes up significantly when a particular item is removed, then the likelihood is that the item is out of place in this test paper. There was no major change to *alpha* for any of the items on the listening paper. In fact the largest positive impact on *alpha* of removing an item was 0.003.

Table 5.3. Mean Facility and Point Biserial (Listening)

Task		Mean Facility	Mean Point Biserial
C&G	1	43.97	0.31
C&G	2	41.95	0.36
C&G	3	32.97	0.38
C&G	4	61.85	0.30
CEsol	5	34.83	0.47
Finn	6	47.04	0.34

In the next set of analyses, we explore any differences in performance across the tasks from the different sources. In the first of the analyses, we compare overall score on the tasks from Communicator and the two Council of Europe standardized tasks.

Table 5.4. shows the descriptive statistics for the tasks from the different sources. All have been averaged to allow for valid comparisons – we do this as the tasks tend to have a different number of items attached. We can see here that there appears to be a difference between the Cambridge items and those of the other two sources (City & Guilds and the Finnish task – note that the four Communicator tasks have been averaged

for this analysis). Table 5.5. confirms that there is a significant effect within the data (i.e. that one of the tasks is significantly more or less difficult than the others), and Table 5.6. confirms that the significant difference lies between the Cambridge ESOL task and the Finnish task, while there is no significant difference between the Communicator task (C&G) and either of the other two tasks. The figures suggest that the Cambridge ESOL task is significantly more difficult than the Finnish task, though Figure 5.4. shows that the City & Guilds tasks and the Finnish tasks are very similar in terms of mean difficulty.

Table 5.4. Descriptive Statistics (Listening)

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Min	Max
					Lower Bound	Upper Bound		
C&G	58	.4519	.15482	.02033	.4112	.4926	.16	.73
CEsol	58	.3483	.31469	.04132	.2655	.4310	.00	1.00
Finn	58	.4704	.23330	.03063	.4091	.5318	.00	.86
Total	174	.4235	.24772	.01878	.3865	.4606	.00	1.00

Table 5.5. One-Way ANOVA for the C&G and other tasks (Listening)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.503	2	.251	4.250	.016
Within Groups	10.114	171	.059		
Total	10.616	173			

Table 5.6. Bonferroni Post hoc Analysis ANOVA for the C&G and other tasks (Listening)

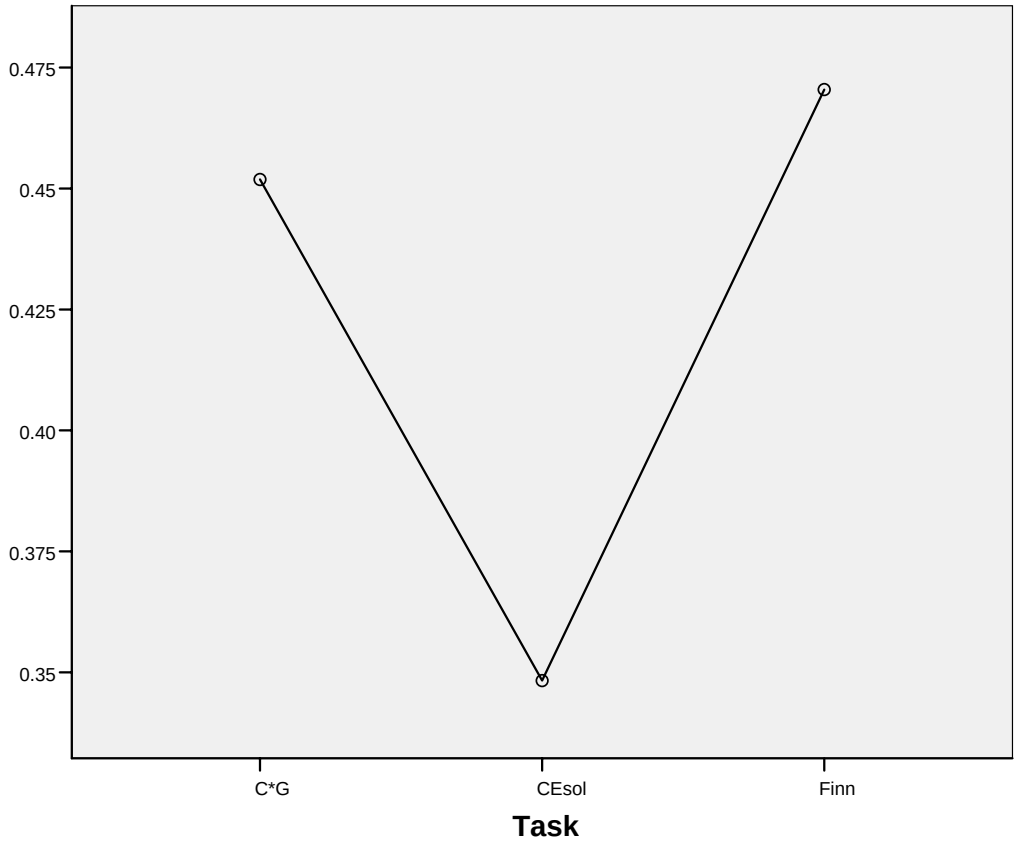
(I) Task	(J) Task	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
C&G	CEsol	.10359	.04516	.069	-.0056	.2128
	Finn	-.01858	.04516	1.000	-.1278	.0906
CEsol	C&G	-.10359	.04516	.069	-.2128	.0056
	Finn	-.12217(*)	.04516	.023	-.2314	-.0130
Finn	C&G	.01858	.04516	1.000	-.0906	.1278
	CEsol	.12217(*)	.04516	.023	.0130	.2314

* The mean difference is significant at the .05 level.

While this should be seen as a satisfactory result from the perspective of this project – it suggests that the Communicator listening paper is at the B2 level as exemplified by the

standardized tasks, there is a real issue here for examination boards who plan to use one of these tasks to establish evidence of the level of their test. It appears that the significant difference in performance on the two tasks may result in examination boards making claims of test level based on evidence that is not as sound as we would wish. The question has to be, "Which of these two tasks (Cambridge ESOL and Finnish) most accurately reflects the B2 level? Clearly, this trial is based on a population that is too small for anything other than doubts to be raised at this point. However, this result concerned us enough to decide that both tasks should be included in the final validation study.

Figure 5.4. Means Plot for the C&G and other tasks (Listening)



Note: the low mean for CEsol indicates that this is more difficult than either of the other two tasks

When a more detailed task level analysis was undertaken, we discovered that one of the Communicator tasks was likely to be at the wrong level (task C&G4). Subsequent exploration revealed that it was a task type for which no changes had been recommended. However, it was also discovered that the particular task used in this trial had been mistakenly included, and had, in fact been rejected for use at the

Communicator level at an earlier time. This required that an alternative live test task of this type had to be located for use in the final validation study.

Table 5.7. Descriptive Statistics for all tasks (Listening)

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Min	Max
					Lower Bound	Upper Bound		
C&G1	58	.4397	.212340	.027882	.38382	.49549	.000	.875
C&G2	58	.4195	.227799	.029911	.35964	.47944	.000	.833
C&G3	58	.3297	.214213	.028128	.27342	.38607	.000	.750
C&G4	58	.6185	.161242	.021172	.57614	.66093	.250	1.000
CEsol	58	.3483	.314694	.041321	.26553	.43102	.000	1.000
Finn	58	.4704	.233298	.030634	.40910	.53179	.000	.857
Total	348	.4377	.248863	.013340	.41146	.46394	.000	1.000

Table 5.8. One-Way ANOVA for all tasks (Listening)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3.118	5	.624	11.608	.000
Within Groups	18.373	342	.054		
Total	21.491	347			

Further analysis of the data (see Tables 5.7. to 5.9.) again indicate a significant effect as we might expect. However, we can see from the post hoc analysis that the main source of this effect is Task4 from the Communicator test. It should also be noted that there is also a significant difference between Task 3 and the Finnish task.

Finally, the means chart (Figure 5.5) confirms the fact that Task 4 is really quite easy when compared to the others, while Task 3 is slightly more difficult than the Cambridge ESOL standardised task.

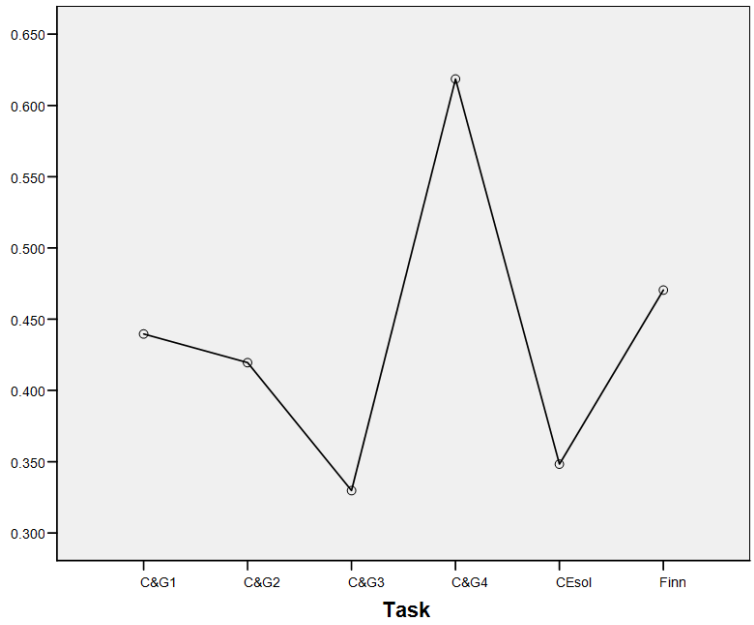
Table 5.9. Bonferroni Post hoc Analysis ANOVA for all tasks (Listening)

(I) Task	(J) Task	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
C&G1	C&G2	.020115	.043040	1.000	-.10711	.14734
	C&G3	.109914	.043040	.166	-.01731	.23714
	C&G4	-.178879(*)	.043040	.001	-.30610	-.05165
	CEsol	.091379	.043040	.517	-.03585	.21860
	Finn	-.030788	.043040	1.000	-.15801	.09644
C&G2	C&G1	-.020115	.043040	1.000	-.14734	.10711
	C&G3	.089799	.043040	.565	-.03743	.21702
	C&G4	-.198994(*)	.043040	.000	-.32622	-.07177
	CEsol	.071264	.043040	1.000	-.05596	.19849
	Finn	-.050903	.043040	1.000	-.17813	.07632
C&G3	C&G1	-.109914	.043040	.166	-.23714	.01731
	C&G2	-.089799	.043040	.565	-.21702	.03743
	C&G4	-.288793(*)	.043040	.000	-.41602	-.16157
	CEsol	-.018534	.043040	1.000	-.14576	.10869
	Finn	-.140702(*)	.043040	.018	-.26793	-.01348
C&G4	C&G1	.178879(*)	.043040	.001	.05165	.30610
	C&G2	.198994(*)	.043040	.000	.07177	.32622
	C&G3	.288793(*)	.043040	.000	.16157	.41602
	CEsol	.270259(*)	.043040	.000	.14303	.39748
	Finn	.148091(*)	.043040	.010	.02087	.27532
CEsol	C&G1	-.091379	.043040	.517	-.21860	.03585
	C&G2	-.071264	.043040	1.000	-.19849	.05596
	C&G3	.018534	.043040	1.000	-.10869	.14576
	C&G4	-.270259(*)	.043040	.000	-.39748	-.14303
	Finn	-.122167	.043040	.072	-.24939	.00506
Finn	C&G1	.030788	.043040	1.000	-.09644	.15801
	C&G2	.050903	.043040	1.000	-.07632	.17813
	C&G3	.140702(*)	.043040	.018	.01348	.26793
	C&G4	-.148091(*)	.043040	.010	-.27532	-.02087
	CEsol	.122167	.043040	.072	-.00506	.24939

* The mean difference is significant at the .05 level.

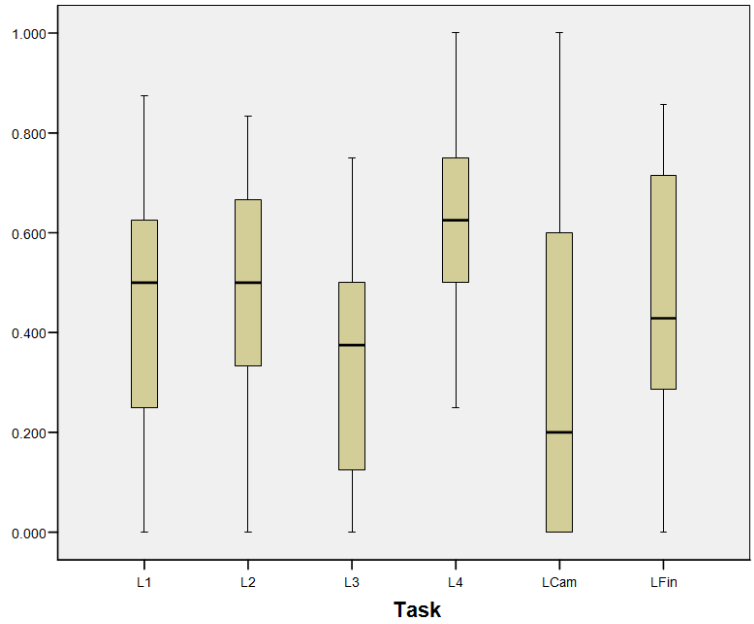
The evidence from this trial suggests that the new and updated tasks are actually working very well and can be seen to be at the level as defined by the standardised tasks. It was anticipated at this stage that the entire listening paper could certainly be shown to be at a level similar to that of the standardised tasks when the correct version of Task 4 is included in the validation study. Therefore, it was decided than any additional work on the listening paper would be very unlikely to add significantly to the evidential basis of our claims.

Figure 5.5. Means Plot for all tasks (Listening)



One final plot shows a potential problem with the trial population and with the Cambridge ESOL task in particular. The Boxplot (Figure 5.6.) shows that all tasks, with the exception of L4 (Communicator Listening Task 4) had at least one instance of a zero score. This may suggest a lack of interest at some point in the trial by a number of participants. We know it is not a single individual as where this happened (with participant 59) we removed all data from the analysis. For the Cambridge ESOL listening task we see the entire range of scores achieved, though here there appears to be a significant proportion of the population who scored zero on this task. Analysis of the raw data indicate that 15 participants (or almost 26%) failed to gain a single point for the items in this task.

Figure 5.6. Box Plot for all tasks (Listening)



5.2.7. Analysis of the Trial Data (Reading)

The mean facility and point biserial estimates for the different tasks are shown in Table 5.10. below. Unlike the listening test, we can see a range of facility levels for the different tasks, with Communicator Task 5 and the Cambridge ESOL task both similarly difficult, and the others generally similar in terms of level. With the reading paper, this is not seen as a problem, as tasks 4 and 5 actually represent two different proposed task forms – the evidence from here is that Task 4 will be included in future papers at this level.

The low mean point biserial found for the Finnish reading task was due to two of the items performing quite poorly. The first of the four items had a facility value of 74.6 and a point biserial of 0.18, while the third item values were 39.0 and -0.1. Since this task was performed late in the process, there may have been a fatigue effect (the Cambridge ESOL task also contained two poorly performing items – #3 with figures of 6.8 and 0.16 and #4 with figures of 18.6 and 0.09).

When the same 'Cronbach's alpha if item deleted' analysis as was used for the listening test data was performed for the reading paper, only one item (Finnish task #3) seemed problematic, though even here the impact on *alpha* would have been 0.008. this suggests that all items were functioning well in distinguishing the stronger from the weaker students.

Table 5.10. Mean Facility and Point Biserial (Reading)

Task		Mean Facility	Mean Point Biserial
C&G	1	34.48	0.38
C&G	2	42.82	0.29
C&G	3	47.13	0.33
C&G	4	40.80	0.38
C&G	5	27.20	0.40
Finn	6	45.69	0.20
CEsol	7	25.00	0.31

Tables 5.11. to 5.13. show a significant effect in the reading data when the mean scores for performance on the tasks from Communicator (C&G) and Cambridge ESOL and Finland are compared. The interesting finding again here is that there appears to be a statistically significant difference in performance between the Cambridge ESOL task and the other two, while the Communicator mean score is similar to that of the Finnish task. This finding is seen very clearly in the means plot (Figure 5.7.).

Table 5.11. Descriptive Statistics (Reading)

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Min	Max
					Lower Bound	Upper Bound		
C&G	58	.3849	.14955	.01964	.3455	.4242	.12	.68
Finn	58	.4569	.22034	.02893	.3990	.5148	.00	1.00
CEsol	58	.2500	.24632	.03234	.1852	.3148	.00	1.00
Total	174	.3639	.22528	.01708	.3302	.3976	.00	1.00

Table 5.12. One-Way ANOVA for the C&G and other tasks (Reading)

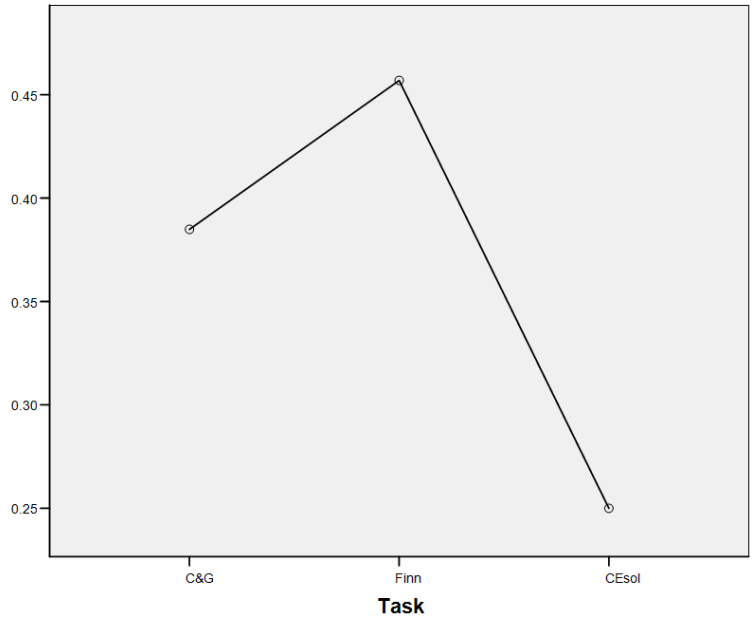
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1.280	2	.640	14.586	.000
Within Groups	7.500	171	.044		
Total	8.780	173			

Table 5.13 Bonferroni Post hoc Analysis ANOVA for the C&G and other tasks (Reading)

(I) Task	(J) Task	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
C&G	Finn	-.07203	.03889	.197	-.1661	.0220
	CEsol	.13487(*)	.03889	.002	.0408	.2289
Finn	C&G	.07203	.03889	.197	-.0220	.1661
	CEsol	.20690(*)	.03889	.000	.1129	.3009
CEsol	C&G	-.13487(*)	.03889	.002	-.2289	-.0408
	Finn	-.20690(*)	.03889	.000	-.3009	-.1129

* The mean difference is significant at the .05 level.

Figure 5.7. Means Plot for the C&G and other tasks (Reading)



As with the listening data, we then analysed all tasks undertaken by the participants. Tables 5.14. to 5.16. which have been summarised in Table 5.17.

Table 5.14. Descriptive Statistics for all tasks (Reading)

Score

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Min	Max
					Lower Bound	Upper Bound		
C&G1	58	.3448	.23124	.03036	.2840	.4056	.00	.83
C&G2	58	.4282	.20969	.02753	.3730	.4833	.00	.83
C&G3	58	.4713	.26332	.03458	.4020	.5405	.00	1.00
C&G4	58	.4080	.17708	.02325	.3615	.4546	.00	.67
C&G5	58	.2720	.16411	.02155	.2289	.3152	.00	.67
Finn	58	.4569	.22034	.02893	.3990	.5148	.00	1.00
CEsol	58	.2500	.24632	.03234	.1852	.3148	.00	1.00
Total	406	.3759	.23191	.01151	.3533	.3985	.00	1.00

Table 5.15. One-Way ANOVA for all tasks (Reading)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2.727	6	.455	9.519	.000
Within Groups	19.054	399	.048		
Total	21.782	405			

Figure 5.8. Means Plot for all tasks (Reading)

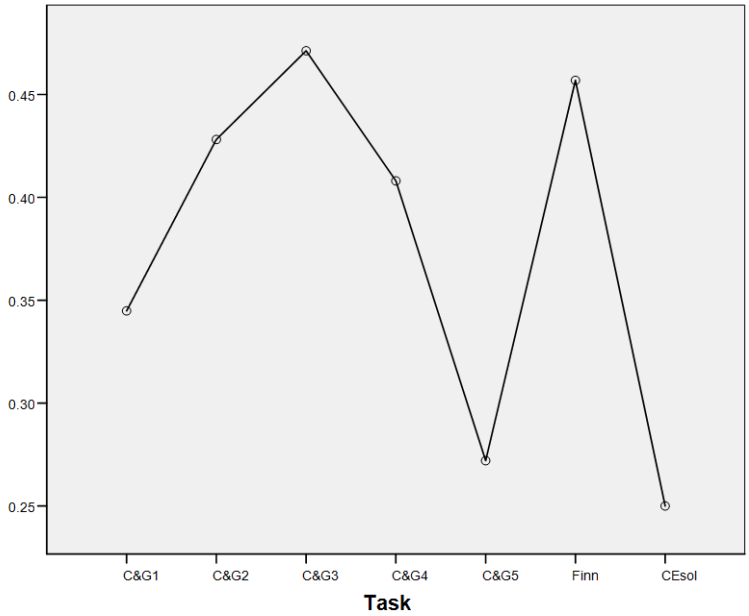


Table 5.16. Bonferroni Post hoc Analysis ANOVA for all tasks (Reading)

(I) Task Id	(J) Task Id	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
C&G1	C&G2	-.08333	.04058	.854	-.2074	.0407
	C&G3	-.12644(*)	.04058	.041	-.2505	-.0024
	C&G4	-.06322	.04058	1.000	-.1873	.0609
	C&G5	.07280	.04058	1.000	-.0513	.1969
	Finn	-.11207	.04058	.126	-.2361	.0120
	CEsol	.09483	.04058	.419	-.0293	.2189
C&G2	C&G1	.08333	.04058	.854	-.0407	.2074
	C&G3	-.04310	.04058	1.000	-.1672	.0810
	C&G4	.02011	.04058	1.000	-.1040	.1442
	C&G5	.15613(*)	.04058	.003	.0321	.2802
	Finn	-.02874	.04058	1.000	-.1528	.0953
	CEsol	.17816(*)	.04058	.000	.0541	.3022
C&G3	C&G1	.12644(*)	.04058	.041	.0024	.2505
	C&G2	.04310	.04058	1.000	-.0810	.1672
	C&G4	.06322	.04058	1.000	-.0609	.1873
	C&G5	.19923(*)	.04058	.000	.0752	.3233
	Finn	.01437	.04058	1.000	-.1097	.1384
	CEsol	.22126(*)	.04058	.000	.0972	.3453
C&G4	C&G1	.06322	.04058	1.000	-.0609	.1873
	C&G2	-.02011	.04058	1.000	-.1442	.1040
	C&G3	-.06322	.04058	1.000	-.1873	.0609
	C&G5	.13602(*)	.04058	.018	.0119	.2601
	Finn	-.04885	.04058	1.000	-.1729	.0752
	CEsol	.15805(*)	.04058	.002	.0340	.2821
C&G5	C&G1	-.07280	.04058	1.000	-.1969	.0513
	C&G2	-.15613(*)	.04058	.003	-.2802	-.0321
	C&G3	-.19923(*)	.04058	.000	-.3233	-.0752
	C&G4	-.13602(*)	.04058	.018	-.2601	-.0119
	Finn	-.18487(*)	.04058	.000	-.3089	-.0608
	CEsol	.02203	.04058	1.000	-.1020	.1461
Finn	C&G1	.11207	.04058	.126	-.0120	.2361
	C&G2	.02874	.04058	1.000	-.0953	.1528
	C&G3	-.01437	.04058	1.000	-.1384	.1097
	C&G4	.04885	.04058	1.000	-.0752	.1729
	C&G5	.18487(*)	.04058	.000	.0608	.3089
	CEsol	.20690(*)	.04058	.000	.0828	.3310
CEsol	C&G1	-.09483	.04058	.419	-.2189	.0293
	C&G2	-.17816(*)	.04058	.000	-.3022	-.0541
	C&G3	-.22126(*)	.04058	.000	-.3453	-.0972
	C&G4	-.15805(*)	.04058	.002	-.2821	-.0340
	C&G5	-.02203	.04058	1.000	-.1461	.1020
	Finn	-.20690(*)	.04058	.000	-.3310	-.0828

* The mean difference is significant at the .05 level.

Table 5.17. Summary of Significant Reading Task Comparisons

	C&G1	C&G2	C&G3	C&G4	C&G5	Finn	CEsol
C&G1	-		●				
C&G2		-			●		●
C&G3	●		-		●		●
C&G4				-	●		●
C&G5		●	●	●	-	●	
Finn					●	-	●
CEsol		●	●	●		●	-

● = significant difference

Note: This table is best read in the same was as a table of correlations (so there is a significant difference between C&G5 and C&G1 for example)

Table 5.17. shows that there are significant differences primarily between all task performances and both Communicator Task 5 and the Cambridge ESOL task. These two are clearly quite different from the others in terms of difficulty, at least for this population.

5.2.8. Analysis of the Trial Data (teacher & self-assessment)

When we compared the teacher and self-assessment data to the score data we found that there was a significant issue with level in both cases. The teachers involved in the project appeared to have significantly inflated perceptions of the level of their students, as the estimations of level were far above the levels suggested by student performances on the various tasks. In the same way, it appears that the data from the Can Do instruments failed to offer any reasonable significant correlations for listening and only an occasional significant correlation of value for reading – note that the highest levels of significance were found in the correlations between the different parts of the Can Do instruments. This confirms our worry that students tended to be very positive in their self-assessment, possibly as a result of their lack of experience with this type of activity.

The 'can do' statements did not work well with this population. Students seem to have over-inflated their level based on the statements and there seems to be little meaningful relationship between the three levels of language described in the 'can do' instrument and performance on any of the tasks. One reason why they are significantly correlated may well be that participants indicated a high level of competence for all items irrespective of level. This suggests that the instrument will need to be changed significantly for the validation study if it is to yield useful results or that the use of this criterion be dropped

from our study. The other option is to ensure that all student who participate in the later parts of the project will first need to receive some training in self-assessment.

In the case of listening, the relatively high (and in all cases significant) correlations between the teacher estimates and the scores on the various tasks performed as well as the self assessments seem to suggest that this was the best predictor of response pattern among the students. It appears that the teachers may have been able to predict the order of ability of their students, though the raw data clearly showed us that they were unable to predict the level. This may be due to a lack of familiarity with the level as described in the CEFR and calls into question the use of such data where the researcher is not certain that the teachers are familiar with the CEFR. The relatively small population in this trial makes further analysis problematic (see Table 5.18).

Table 5.18 Correlations of Variables for Listening

		AveL1	AveL2	AveL3	AveL4	CEsol	Finn	B1	B2	C1	T'ch
AveL1	Pearson	1	.449(**)	.409(**)	.397(**)	.359(**)	.254	.004	.220	.095	.447(**)
	Sig.		.000	.001	.002	.006	.054	.979	.097	.478	.000
AveL2	Pearson	.449(**)	1	.448(**)	.433(**)	.398(**)	.387(**)	.178	.206	.257	.448(**)
	Sig.	.000		.000	.001	.002	.003	.182	.121	.051	.000
AveL3	Pearson	.409(**)	.448(**)	1	.452(**)	.401(**)	.383(**)	.169	.221	.227	.392(**)
	Sig.	.001	.000		.000	.002	.003	.204	.096	.087	.002
AveL4	Pearson	.397(**)	.433(**)	.452(**)	1	.356(**)	.357(**)	.048	.224	.209	.345(**)
	Sig.	.002	.001	.000		.006	.006	.722	.091	.116	.008
CEsol	Pearson	.359(**)	.398(**)	.401(**)	.356(**)	1	.460(**)	.070	.081	.137	.283(*)
	Sig.	.006	.002	.002	.006		.000	.601	.546	.307	.031
Finn	Pearson	.254	.387(**)	.383(**)	.357(**)	.460(**)	1	.081	.151	.228	.413(**)
	Sig.	.054	.003	.003	.006	.000		.548	.258	.085	.001
B1	Pearson	.004	.178	.169	.048	.070	.081	1	.644(**)	.652(**)	.306(*)
	Sig.	.979	.182	.204	.722	.601	.548		.000	.000	.019
B2	Pearson	.220	.206	.221	.224	.081	.151	.644(**)	1	.685(**)	.361(**)
	Sig.	.097	.121	.096	.091	.546	.258	.000		.000	.005
C1	Pearson	.095	.257	.227	.209	.137	.228	.652(**)	.685(**)	1	.388(**)
	Sig.	.478	.051	.087	.116	.307	.085	.000	.000		.003
T'ch	Pearson	.447(**)	.448(**)	.392(**)	.345(**)	.283(*)	.413(**)	.306(*)	.361(**)	.388(**)	1
	Sig.	.000	.000	.002	.008	.031	.001	.019	.005	.003	

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

Note: In both this table and Table 5.19 the following abbraviations are used:

Ave1 = Mean for C&G Task 1	CEsol = Mean for Cambridge ESOL Task
Ave2 = Mean for C&G Task 2	Finn = Mean for Finnish Task
Ave3 = Mean for C&G Task 3	B1 = Can Do statements aimed at level B1
Ave4 = Mean for C&G Task 4	B2 = Can Do statements aimed at level B2
T'ch = Teacher estimates	C1 = Can Do statements aimed at level C1

The situation with reading seems to be somewhat different (Table 5.19). The teacher predictions in this case seem to be correlated significantly with fewer of the other variables. We can also see that the scores for the Finnish task do not seem to correlate well with the other variables, though the scores for the Finnish listening task seem to correlate well with those of the other listening variables.

Table 5.19. Correlations of Variables for Reading

		AveR1	AveR2	AveR3	AveR4	AveR5	Finn	CEsol	B1	B2	C1	T'ch
AveR1	Pears'n	1	.510(**)	.294(*)	.423(**)	.456(**)	.239	.376(**)	.283(*)	.360(**)	.202	.371(**)
	Sig.		.000	.025	.001	.000	.070	.004	.031	.005	.129	.004
AveR2	Pears'n	.510(**)	1	.333(*)	.375(**)	.493(**)	.153	.109	.309(*)	.277(*)	.064	.209
	Sig.	.000		.011	.004	.000	.251	.418	.018	.036	.633	.116
AveR3	Pears'n	.294(*)	.333(*)	1	.339(**)	.274(*)	.192	.225	.193	.162	.093	.204
	Sig.	.025	.011		.009	.037	.148	.089	.146	.224	.488	.125
AveR4	Pears'n	.423(**)	.375(**)	.339(**)	1	.444(**)	.022	.272(*)	.321(*)	.197	.072	.306(*)
	Sig.	.001	.004	.009		.000	.873	.039	.014	.139	.592	.019
AveR5	Pears'n	.456(**)	.493(**)	.274(*)	.444(**)	1	.263(*)	.129	.302(*)	.291(*)	.089	.372(**)
	Sig.	.000	.000	.037	.000		.046	.336	.021	.027	.508	.004
Finn	Pears'n	.239	.153	.192	.022	.263(*)	1	.148	.097	.277(*)	-.027	.305(*)
	Sig.	.070	.251	.148	.873	.046		.267	.467	.035	.840	.020
CEsol	Pears'n	.376(**)	.109	.225	.272(*)	.129	.148	1	-.096	.239	.064	.164
	Sig.	.004	.418	.089	.039	.336	.267		.472	.071	.631	.220
B1	Pears'n	.283(*)	.309(*)	.193	.321(*)	.302(*)	.097	-.096	1	.647(**)	.708(**)	.425(**)
	Sig.	.031	.018	.146	.014	.021	.467	.472		.000	.000	.001
B2	Pears'n	.360(**)	.277(*)	.162	.197	.291(*)	.277(*)	.239	.647(**)	1	.596(**)	.531(**)
	Sig.	.005	.036	.224	.139	.027	.035	.071	.000		.000	.000
C1	Pears'n	.202	.064	.093	.072	.089	-.027	.064	.708(**)	.596(**)	1	.256
	Sig.	.129	.633	.488	.592	.508	.840	.631	.000	.000		.053
T'ch	Pears'n	.371(**)	.209	.204	.306(*)	.372(**)	.305(*)	.164	.425(**)	.531(**)	.256	1
	Sig.	.004	.116	.125	.019	.004	.020	.220	.001	.000	.053	

5.3. The Expert Panel (Standard Setting)

The expert panels for the standard setting events also contained a mix of insiders – the original review panel members plus a number of trained and experienced B2 level item/task writers – and three outsiders, two of the original group plus one very highly experienced person with recent experience of a CEFR linking project and many years of experience as a test developer. Finally, CLARe again provided an independent chair for the sessions.

The three expert panel events are reported on in the following three sub-sections.

5.3.1. The First Panel Event (Reading)

The standard setting approach chosen for the receptive skills was the extended Angoff. This was partly because "the Angoff method appears to offer the best balance between technical adequacy and practicability" (Berk, 1986: 147) and partly because this variation has replaced the original (generally modified) version used for tests comprised predominantly of multiple choice items (Cisek and Bunch, 2007: 82).

The procedure adopted for this event was in two stages:

Prior to the Event

1. Re-familiarisation of judges with the CEFR levels B1, B2 and C1 – the primary focus being on level B2.
2. Familiarisation of judges with the test tasks.
3. Pre-event estimation of likelihood of minimally competent candidate at B2 answering each item correct (Yes/No – coded as 1 and 0, a variation based on Impara & Plake, 1998) together with an estimate of how certain the judge is of this decision. This final element was only used to inform the discussions and was not included in the calculation of the cut scores. This was because we felt that the indication of certainty was more valuable as a stimulant to discussion to help judges to consider their decisions. By the time the second round of judging came around judges tended to be more clear as to why they made the decisions they did and the MFR analysis confirms that they were internally consistent, thus eliminating the need to use these figures.

During the Event

1. Clarification and finalization of a definition of the minimally competent candidate at level B2, based on a preliminary definition developed by the project team prior to the event.
2. Discussion of pre-event judgements, in three parts: a) review of the tasks and the decisions made, b) discussion based on presentation of preliminary analysis of the data from these judgements, c) discussion based on item statistics from the administration of the tasks to the main validation population. We made a change to the Manual recommendations at this point, deciding to include the Council of Europe recommended tasks in the level estimation task rather than use them for pre-event training. This was done as the participants were all experienced in

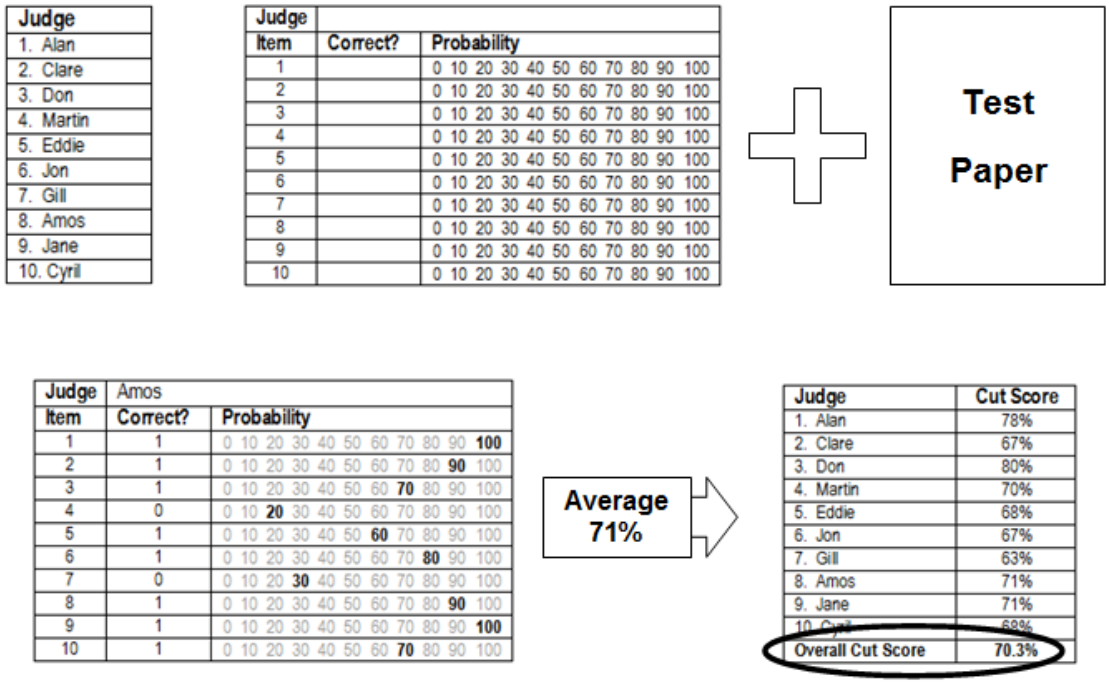
teaching and testing at the level and had already been through an extensive training schedule prior to this project.

3. Round 2 of judgements, with judges asked to take the previous discussions into consideration.

4. Estimation of the cut-score based on a multi-faceted Rasch analysis of the data from Round 2 of the judgements.

The process used for the reading and listening standardisation events is exemplified in Figure 5.9. Having first selected the panel members, we ask each one to review our test paper and to complete a response sheet like the one shown. On this sheet, the judge indicates whether the minimally competent learner will answer each item in the paper correctly (1) or incorrectly (0). They then indicate the percentage of minimally competent learners who will answer each item correctly. The complete form from one of the judges (Amos) is shown next. To calculate Amos' indicative cut score we average the Probability scores (in his case the average is 71%). If we then take this calculation for all of the judges and find the average, we come to the suggested cut score for the group.

Figure 5.9. Example of the Angoff procedure used in this project



Preliminary Decisions

It was noted that a number of the judges in their pre-event submission were particularly high in their estimation of the number of items a minimally competent candidate at level B2 would answer correctly. This indicated that there was some confusion about the actual decision the group had been asked to make. When the event began, we decided to address this situation by focusing more clearly on an agreed definition of this candidate.

Based on this discussion, the definition of the minimally competent candidate at this level, adapted from the Overall Reading Comprehension Scales (CEFR, 2001: 69) and the Reading for Information and Argument Scale (CEFR, 2001: 70) was agreed on. The definition was:

> Can understand without dependence on dictionaries or glossaries articles, reports and narratives aimed at the general reader and texts in which the writers adopt particular stances or viewpoints. Has difficulty with specialized or unclearly structured texts and low frequency lexis.

Outcomes

Following a further discussion of the purpose of this event (to identify the least able candidate who should be awarded a passing grade on the Communicator examination based on our interpretation of the CEFR Level B2) we decided to ask the judges to repeat their evaluation of the items. Again, judges were asked to say whether the minimally competent candidate would answer each item correctly (1) or incorrectly (0) and also asked to indicate how confident they felt about each judgement. This latter variable was then used as an intervening variable in a multi-faceted analysis of the judgements made.

The results of this phase of the event indicated that two of the judges were still quite a bit away from the rest of the group and the decision was made at this point to drop their estimates from the analysis. The data were then re-analyzed using the nine remaining judges. The results (Table 5.20) indicate that all of the remaining judges were both very close in terms of agreement and all were all internally consistent.

In addition to this finding, it is clear from Table 5.21. that all of the items were relatively easy to assign decisions to: note the range of infit mean square estimates – all within the range of 0.5 to 1.5 adopted by Lunz & Wright (1997: 83). We therefore feel that we can support any decisions suggested by this group.

Table 5.20. MRF Judge Measurement Table (Round 2 – Reading)

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Nu	Judges
16	29	.6	.58	-.19	.41	1.0	0	.9	0	1	CG1
18	29	.6	.66	-.52	.44	1.1	0	1.4	1	4	CG4
14	29	.5	.47	.25	.42	1.1	0	1.1	0	5	Ou1
16	29	.6	.53	.02	.42	.9	0	.8	0	6	CG5
18	29	.6	.60	-.27	.44	1.3	1	1.3	0	7	CG6
13	29	.4	.43	.41	.42	.9	0	.8	0	8	CG7
14	29	.5	.44	.39	.41	.9	0	1.0	0	9	CG8
18	29	.6	.60	-.27	.43	.7	-1	.7	-1	10	Ou2
15	29	.5	.49	.17	.42	1.0	0	1.1	0	11	Ou3
15.8	29.0	.5	.53	.00	.42	1.0	-.1	1.0	-.1	Mean (Count: 9)	
1.8	.0	.1	.08	.31	.01	.1	.8	.2	.8	S.D.	

RMSE (Model) .42 Adj S.D. .00 Separation .00 Reliability .00
Fixed (all same) chi-square: 4.7 d.f.: 8 significance: .78

Table 5.21. MRF Item Measurement Table (Round 2 – Reading)

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Nu	Items
4	9	.4	.37	-.54	.70	.9	0	.8	0	1	1.1
5	9	.6	.51	.04	.70	.9	0	.9	0	2	1.2
7	9	.8	.76	1.15	.83	1.0	0	1.1	0	3	1.3
8	9	.9	.87	1.89	1.07	1.1	0	1.2	0	4	1.4
6	9	.7	.69	.79	.75	1.1	0	1.0	0	5	1.5
6	9	.7	.65	.60	.74	1.2	0	1.1	0	6	1.6
5	9	.6	.57	.28	.70	1.3	1	1.3	1	7	2.1
4	9	.4	.36	-.59	.68	1.1	0	1.1	0	8	2.2
4	9	.4	.45	-.21	.70	1.0	0	1.0	0	9	2.3
2	9	.2	.21	-1.31	.83	.9	0	.8	0	10	2.4
2	9	.2	.20	-1.37	.81	1.0	0	1.0	0	11	2.5
0	9			(-3.62	1.86)	Minimum				12	2.6
2	9	.2	.17	-1.59	.81	1.0	0	1.0	0	13	3.2
5	9	.6	.53	.12	.70	.8	0	.8	0	14	3.3
7	9	.8	.78	1.29	.83	.9	0	.9	0	15	3.4
5	9	.6	.53	.10	.69	1.0	0	1.0	0	16	3.5
7	9	.8	.82	1.52	.84	1.0	0	1.2	0	17	3.6
6	9	.7	.65	.61	.71	1.1	0	1.1	0	18	3.7
8	9	.9	.88	1.95	1.06	1.1	0	1.5	0	19	3.8
2	9	.2	.17	-1.61	.81	1.2	0	1.4	0	20	3.9
6	9	.7	.67	.71	.73	1.3	1	1.3	0	21	3.1
4	9	.4	.43	-.29	.68	1.0	0	1.0	0	22	4.1
4	9	.4	.45	-.19	.68	1.0	0	1.0	0	23	4.2
6	9	.7	.65	.63	.73	.8	0	.8	0	24	4.3
2	9	.2	.23	-1.20	.82	.9	0	.8	0	25	4.4
7	9	.8	.79	1.34	.82	.8	0	.7	0	26	4.5
6	9	.7	.72	.92	.75	.8	0	.7	0	27	4.6
2	9	.2	.18	-1.53	.81	.9	0	.8	0	28	4.7
5	9	.6	.60	.40	.69	.9	0	.9	0	29	4.8
5	9	.6	.53	.13	.70	.9	0	.9	0	30	4.9
4.7	9.0	.5	.51	.14	.77	1.0	.0	1.0	-.1	Mean (Count: 30)	
2.0	.0	.2	.23	1.03	.10	.1	.5	.2	.5	S.D.	

RMSE (Model) .78 Adj S.D. .67 Separation .86 Reliability .43
Fixed (all same) chi-square: 43.3 d.f.: 28 significance: .03

This table also indicates that the average of the fair average scores for all items is .514. This suggests that the cut score should be set at 15.30. For obvious reasons, this cannot be used operationally, so based on further discussion within the group, the actual cut score was finally set at 15. The reliability of the judges was estimated using Cronbach's Alpha and was found to be 0.850 – inter-class correlation was also calculated as it represents a more valid estimate for these circumstances and was found to be 0.848. The figures presented here indicate that the level of agreement and consistency of the judges was acceptably high.

Commentary

The standard setting event proved successful in that a final cut score was agreed. This cut score was in line with current practice at City & Guilds, and so the approach within the organization was also supported. The decision to include both internal and external judges was certainly vindicated by the level and outcome of the various discussions, and clearly adds weight to the claim that the agreed cut score is linked to the CEFR Level B2.

It would also appear that the decision to use the extended Angoff approach was justified, though there is considerable doubt that this approach would have generated the kind of discussion-led decisions had the make-up of the group been different. We feel that a group comprised only of insiders or outsiders would not offer the same balance of knowledge of the test itself (not simply the test tasks or items being reviewed, but knowing that these represent typical test tasks or items and also knowing the process behind the development, writing and operational value of these tasks and items) with a broader awareness of language learning and assessment at the level under consideration (and the other level which border it) and a keen operational understanding of the CEFR and its role in learning and assessment.

It is also clear to us that our decision to undertake a preliminary critical review of the Communicator test, which resulted in some small changes to the test itself, not necessarily in terms of level but a tightening up of the specifications and of the actual tasks and items, was justified. In fact, we firmly believe that without such a critical review there is a real danger that any linking project will be likely to either prove meaningless or will fail at the final validation phase – particularly if we look beyond the primarily psychometric model of validation that the pilot manual seems to promote to a broader validation report based on a model such as that provided in Weir's frameworks (2005).

5.3.2. The Second Panel Event (Listening)

The same standard setting approach that had been chosen for the reading event was also used for the listening.

The procedure adopted for this event was also in two stages:

Prior to the Event

1. Re-familiarisation of judges with the CEFR levels B1, B2 and C1 – the primary focus being on level B2.
2. Familiarisation of judges with the test tasks.

3. Pre-event estimation of likelihood of minimally competent candidate at B2 answering each item correct (Yes/No – coded as 1 and 0) together with an estimate of how certain the judge is of this decision. The certainty estimates were used here in the same way as they were used for the Reading, see above).

During the Event

1. Clarification and finalization of a definition of the minimally competent candidate at level B2, based on a preliminary definition developed by the project team prior to the event (the same approach as was taken for the Reading).
2. Discussion of pre-event judgements, in three parts: a) review of the tasks and the decisions made, b) discussion based on presentation of preliminary analysis of the data from these judgements, c) discussion based on item statistics from the administration of the tasks to the main validation population.
3. Round 2 of judgements, with judges asked to take the previous discussions into consideration.
4. Estimation of the cut-score based on a multi-faceted Rasch analysis of the data from Round 2 of the judgements.

Preliminary Decisions

Based on the discussions which took place during the early stages of the event, the definition of the least able candidate at this level was defined as:

> Can follow most standard spoken language, live or broadcast, such as lectures, discussion and debates, on topics normally encountered in personal, social, academic or vocational life. Has difficulty understanding implicit meaning in extended speech and finds it difficult to understand if there is extreme background noise, inadequate discourse structure and idiomatic usage.

Outcomes

As happened during the Reading Paper standard setting event, we asked the judges to repeat their evaluation of the items following a lot of discussion and re-listening and assessment of the items. Again, judges were asked to say whether the least capable candidate would answer each item correctly (1) or incorrectly (0) and also asked to indicate how confident they felt about each judgement. This latter variable was then used

as an intervening variable in a multi-faceted analysis of the judgements made (as had been the case for the Reading Paper.

Table 5.22. MRF Judge Measurement Table (Round 2 – Listening)

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Nu	Judges
14	28	.5	.57	-.34	.53	1.0	0	.7	0	1	CG1
18	30	.6	.63	-.60	.52	.8	0	.6	0	2	CG2
16	30	.5	.28	.85	.47	1.1	0	.9	0	3	CG3
16	30	.5	.53	-.21	.51	1.1	0	.9	0	4	CG4
16	30	.5	.17	1.51	.48	.7	-1	.5	0	5	Ou1
14	29	.5	.54	-.22	.53	1.0	0	1.5	0	6	CG5
16	30	.5	.58	-.40	.48	1.2	0	.9	0	7	CG6
18	30	.6	.25	1.00	.54	.9	0	.6	0	8	CG7
16	30	.5	.34	.61	.49	.8	-1	.5	0	9	CG8
25	30	.8	.84	-1.75	.66	1.4	1	1.0	0	10	Ou2
22	30	.7	.59	-.46	.60	1.3	0	1.0	0	11	Ou3
17.4	29.7	.6	.48	.00	.53	1.0	.0	.8	-.2	Mean (Count: 11)	
3.2	.6	.1	.19	.87	.05	.2	.8	.3	.3	S.D.	

RMSE (Model) .53 Adj S.D. .69 Separation 1.30 Reliability .63
Fixed (all same) chi-square: 27.8 d.f.: 10 significance: .00

Table 5.22. indicates that the judges were all internally consistent in their judgements, note the range of infit mean square estimates – all within the range of 0.5 to 1.5 adopted by Lunz & Wright (1997: 83). We therefore feel that we can support any decisions suggested by this group.

This is not to say that there is perfect agreement between the judges. Clearly this is not the case (note the all same chi-square result (rejecting the null hypothesis that the group is fixed or 'all same) which indicates that the group is not homogenous in terms of their estimation of the cut score). The advantage to using MFR is clear at this point as it allow us to take into account the variation in judgements to make a fair average estimate of item difficulty – and so bolster the validity of the suggested cut-score.

Table 5.23 indicates that the items ranged in terms of how difficult the judges found them to categorise (whether they could be answered correctly or not by a least able candidate at CEFR Level B2) – see the range of infit mean square estimates. The fair average difficulty estimate, which took into account the degree to which each judge was certain of his/her decision, is .49. When related to the number of items included in the paper used for the standard setting event (N = 30) we find that the cut score should be set at 14.70. For operational reasons (again we need a whole number as the cut scores will be based on raw score data since this is an on-demand examination) the cut score recommended from this event is 15.

Table 5.23. MRF Item Measurement Table (Round 2 – Listening)

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Nu	Items
9	10	.9	.96	3.22	1.12	.8	0	.4	0	1	1.1.1
9	11	.8	.91	2.32	1.09	1.4	0	.6	0	2	1.1.2
2	11	.2	.04	-3.06	.97	.7	0	.4	0	3	1.2.1
3	11	.3	.08	-2.50	.81	1.1	0	1.0	0	4	1.2.2
8	11	.7	.79	1.32	.86	.4	-1	.3	0	5	1.3.1
7	11	.6	.57	.27	.74	.9	0	.6	0	6	1.3.2
6	11	.5	.63	.54	.73	1.4	1	1.4	0	7	1.4.1
10	11	.9	.76	1.13	1.08	.9	0	.5	0	8	1.4.2
8	11	.7	.71	.90	.88	.6	-1	.3	0	9	2.1.1
2	10	.2	.08	-2.41	1.07	2.2	1	2.0	0	10	2.1.2
8	11	.7	.44	-.22	.78	.7	0	.5	0	11	2.2.1
5	11	.5	.22	-1.26	.81	1.2	0	1.3	0	12	2.2.2
5	11	.5	.17	-1.58	.73	1.7	1	1.9	0	13	2.3.1
2	11	.2	.07	-2.59	1.03	.4	-1	.2	0	14	2.3.2
8	11	.7	.47	-.12	.98	.6	0	.3	0	15	3.1
7	11	.6	.43	-.30	.90	1.1	0	.6	0	16	3.2
6	11	.5	.47	-.11	.71	1.1	0	1.2	0	17	3.3
4	10	.4	.13	-1.92	.78	1.1	0	1.0	0	18	3.4
7	11	.6	.57	.28	.84	1.0	0	.6	0	19	3.5
3	11	.3	.09	-2.30	1.02	2.8	1	2.2	0	20	3.6
7	11	.6	.54	.16	.79	.9	0	.6	0	21	3.7
7	11	.6	.49	-.03	.74	1.0	0	.9	0	22	3.8
7	11	.6	.61	.46	.81	.4	-1	.3	-1	23	4.1
9	11	.8	.93	2.64	.96	.6	0	.3	0	24	4.2
7	11	.6	.41	-.38	.82	1.1	0	.9	0	25	4.3
10	11	.9	.89	2.05	1.27	.4	-1	.1	0	26	4.4
6	11	.5	.63	.52	.66	.9	0	.8	0	27	4.5
6	11	.5	.53	.13	.93	.3	-1	.2	-1	28	4.6
10	11	.9	.95	3.04	1.13	1.5	0	2.9	0	29	4.7
3	11	.3	.08	-2.44	.84	.6	-1	.4	0	30	4.8
6.4	10.9	.6	.49	-.07	.90	1.0	-.2	.8	-.2	Mean (Count: 30)	
2.4	.3	.2	.29	1.71	.15	.5	.9	.7	.5	S.D.	

RMSE (Model) .91 Adj S.D. 1.45 Separation 1.60 Reliability .72
Fixed (all same) chi-square: 92.8 d.f.: 29 significance: .00

Commentary

The standard setting event again proved successful in that a final cut score was agreed. Like with the Reading Paper, this cut score was in line with current practice at City & Guilds, and so the approach within the organization was again supported. However, the practice of asking panel members to indicate their certainty in terms of the decisions they made was reconsidered, as it became apparent that the uncertainty could result in a decision going either up or down. The considerable discussions that followed the judging rounds was at least in part due to this lack of clarity, so for this reason we decided to change our approach in the later projects to one in which panel members would indicate a probability estimate for each item.

5.3.3. The Third Panel Event (Writing)

The third and final panel event was set in place to establish that the decisions made in the assessment of writing in the Communicator examination could be seen to be in line with the descriptions of the level described at CEFR B2. The initial critical review of the examination indicated that the tasks were likely to result in performances at B2. However, there was some concern expressed with the potential for decisions to be made

regarding those performances that might overestimate the level of the candidates. For this reason it was recommended at the time that City & Guilds review their recruitment and training procedures for test writers and examiners to include a more explicit role for the CEFR in the process. This recommendation was acted upon by the institution and the standard setting event was then based on the situation that prevailed following the changes.

The focus of the standard setting event was somewhat different to those for the reading and listening events. In those events, the judges were asked to make judgements based on their interpretation of the test items in order to set a meaningful (in terms of minimally competent candidate at CEFR Level B2) cut score. In this event, the judges were to look at actual task performances and to make similar judgements. In order to assess the quality of the decisions made, we decided to include a number of additional sample performances that had been suggested by the Council of Europe to be representative of the level. In fact, as we were keen to look beyond Level B2 and to include tasks at levels B1 and C1 in order to ensure that the focus level was being accurately judged.

The procedure adopted for this event was, like the previous events, also in two stages:

Prior to the Event

1. Re-familiarisation of judges with the CEFR levels B1, B2 and C1 – the primary focus being on level B2.
2. Familiarisation of judges with the test tasks and Council of Europe standardized tasks.
3. Pre-event judgements of how the different tasks should be scored using the City & Guilds rating scale for writing, which had been developed based on the descriptors of writing at level B2 in the CEFR. Judges were asked to make one of a number of decisions:
    - 0 – clearly below level B2
    - 1 – Fail at level B2
    - 2 – Pass at level B2
    - 3 – First Class Pass at level B2
    - 4 – clearly above level B2

During the Event

1. Discussion of pre-event judgements, in three parts: a) review of the tasks and the decisions made, b) discussion based on presentation of preliminary analysis of the data from these judgements.
2. Round 2 of judgements, with judges asked to take the previous discussions into consideration.
3. Estimation of the equivalence of the tasks included in the event based on a multi-faceted Rasch analysis of the data from Round 2 of the judgements.

Preliminary Decisions

Based on the pre-event judgements and the initial discussions it became clear that there were relatively few differences between the judges. At this point the lack of a range of Council of Europe standardized tasks was noted by the judges as being a problem – it was felt that a single task to define each level was likely to be problematic. This is particularly the case when we look to the outcomes of the trials of the updated Communicator tasks reported on earlier (see Table 5.10) and in the validation study, reported on in the following part of this report, in which it appears that there are different levels of ability even among the so-called standardized tasks (we would argue that these tasks have not been standardized as there is no evidence that they are at the same level except for the basic descriptions presented by the developers of those tasks – until there is empirical evidence linking the performances they should only be referred to as tasks that have been claimed to be indicative of the level).

Outcomes

The tasks were coded for analysis, though these codes were not made known to the judges until after the event. The codes, See Table 5.24. indicate the origin (C&G = City and Guilds; Camb = Cambridge ESOL), the intended level (B1, B2 and C1) and the actual test score awarded where applicable (for City & Guilds tasks these were F = fail; P = pass; FP = first class pass).

After the second round of judging, the MFR analysis was carried out. Judges had been asked to indicate the level at which they felt each task was aimed and this was used as an intervening variable in the final analysis. In addition, since the judges were asked to use the three criteria (Range, Accuracy and Organisation) to inform their final judgement, the decisions made for these three criteria plus the actual final decision were all included in

the analysis. It should be noted at this stage that additional analyses which eliminated the estimate of task level, included only the criteria decisions and analysed only the final decision were all carried out. No significant difference was found between all of these analyses so the original one is reported here.

Table 5.24. Task Performance Coding and Expected Level (Writing)

#	Code	Expected
1	1C&GB2F	1
2	2C&GB2FP	3
3	3C&GB2P	2
4	4C&GB1FP	3
5	5C&GC1P	4
6	6C&GB1P	0
8	8CambB1	0
9	9CambB2	2-3
10	10C&GB1P	0
11	11CambC1	4
12	12C&GB2F	0-1
13	13C&GB2FP	3
14	14C&GC1F	0-3
15	15C&GB1FP	0-1

Note: For the City & Guilds tasks the following suffix codes apply:

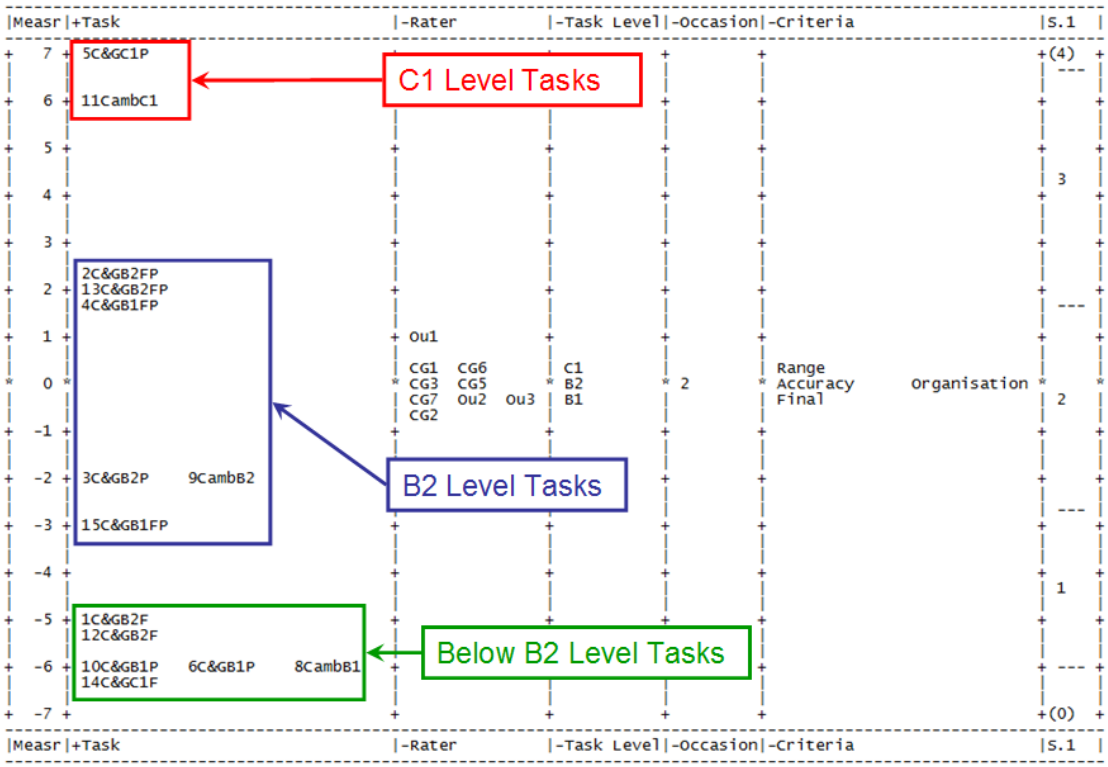
P = Pass at the level indicated before this suffix

FP = First Class Pass at the level indicated before this suffix

F = Fail at the level indicated before this suffix

Figure 5.10. indicates that the analysis found there to be three distinct groups of task performances, which we have interpreted as indicating the levels C1, B2 and B1. Since we were focused on level B2, we should not attempt to over-interpret the other level here – remember that the judges were asked to place the performances in relation to a B2 level rating scale and were not told that there were performances that were actually from the other levels.

Figure 5.10. MRF All Facet Rulers – Summary Chary (Round 2 – Writing)



Some of the interesting things to emerge were:

- The strong level of agreement between the original scores awarded and the levels indicated by the judges.
- There were a few rogue performances, but these can be explained relatively easily. Item 15C&GB1FP was seen by the judges to be at Level B2, even though it was a B1 examination performance. The fact that it was originally awarded a First Class Pass can be seen as an indication that this award is right on the border with the next level up – a situation that is also clear from items 2C&GB2FP, 4C&GB2FP and 13C&GB2FP, and also with items 1C&GB2F and 12C&GB2F which were seen to be Level B2 fails but could also be seen as being located at Level B1. The strangest item is 14C&GC1F, which was originally a failure at Level C1 and might have been expected to fall into the B2 zone. However, it was seen by these judges as being below Level B2 altogether. This is of course, plausible, though is does raise some question about the value in including such a performance in a standard setting event such as this.
- The perception of the judges that the First Class Pass performances at B2 were seen to be quite separate from the lower Pass level performances, though there is

a clear gap between these three and the two City & Guilds tasks (1C&GB2F and 12C&GB2F) that were adjudged to have failed to reach the standard of B2.

The above findings are confirmed by the data in Table 5.25., which show that the performances range across the expected levels. Further analysis of these detailed results can be found in Table 2.26. indicates that the judges agreed with the original expected level on 11 of the 14 cases (note that there was no Task 7 – the numbering of the tasks was not consecutive due to an administrative error). Of the remaining three tasks, the rounding exaggerated the difference on two occasions, where there was very little real difference between the judgements and the expected level. On one occasion (11CambC1) the difference was 0.7 of a level. This may have been due, at least in part, to the reluctance of the judges to award level 4. This appears to be confirmed by the probability curves (Figure 2.10.), which show a slightly greater tendency towards awarding a 3 than might be expected.

Table 5.25. MRF Performance Measurement Table (Round 2 – Writing)

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Nu	Task
30	36	.8	.83	-4.84	.31	.6	-1	.6	-1	1	1C&GB2F
97	36	2.7	2.69	2.50	.35	1.0	0	1.0	0	2	2C&GB2FP
58	36	1.6	1.63	-2.11	.32	.9	0	.9	0	3	3C&GB2P
91	36	2.5	2.48	1.68	.33	1.0	0	1.0	0	4	4C&GB1FP
127	36	3.5	3.54	6.95	.34	1.2	1	1.2	1	5	5C&GC1P
20	36	.6	.51	-5.96	.32	1.0	0	1.0	0	6	6C&GB1P
18	36	.5	.46	-6.13	.32	1.4	1	1.3	1	8	8CambB1
60	36	1.7	1.67	-1.96	.33	.7	-1	.7	-1	9	9CambB2
19	36	.5	.47	-6.10	.32	.9	0	.9	0	10	10C&GB1P
119	36	3.3	3.30	5.99	.35	.8	-1	.7	-1	11	11CambC1
24	36	.7	.66	-5.42	.31	1.0	0	1.0	0	12	12C&GB2F
93	36	2.6	2.58	2.07	.33	.9	0	.9	0	13	13C&GB2FP
14	36	.4	.37	-6.47	.34	1.3	1	1.2	1	14	14C&GC1F
49	36	1.4	1.36	-3.01	.31	.9	0	.9	0	15	15C&GB1FP
58.5	36.0	1.6	1.61	-1.63	.33	1.0	-.1	1.0	-.2	Mean (Count: 14)	
38.6	.0	1.1	1.08	4.49	.01	.2	1.2	.2	1.1	S.D.	

RMSE (Model) .33 Adj S.D. 4.48 Separation 13.69 Reliability .99
Fixed (all same) chi-square: 2521.6 d.f.: 13 significance: .00

We can also see from Table 5.22. that the performances were all relatively stable in terms of how the judges were able to award levels. The infit mean square estimates are all quite low, with no sample performance suggesting inconsistency (note: the same range of acceptability was used here as was used for the listening and reading, so any infit mean square estimate that is under 1.5 is seen to be satisfactory). The very small changes between the observed average (i.e. the mathematical average of the levels designated) and the fair average (the average which takes into account the data related to the other variables (expected level, criteria, judge harshness), also confirm the high level of agreement between the judges.

Table 5.26. Item Expected and Final Level (Writing)

Item	Expected	Final	Rounded	Agreement
1C&GB2F	1	.83	1	Yes
2C&GB2FP	3	2.69	3	Yes
3C&GB2P	2	1.63	2	Yes
4C&GB1FP	3	2.48	2	No (slight under)
5C&GC1P	4	3.54	4	Yes
6C&GB1P	0	.51	1	No (slight over)
8CambB1	0	.46	0	Yes
9CambB2	2-3	1.67	2	Yes
10C&GB1P	0	.47	0	Yes
11CambC1	4	3.3	3	No (under)
12C&GB2F	0-1	.66	1	Yes
13C&GB2FP	3	2.58	3	Yes
14C&GC1F	0-3	.37	0	Yes
15C&GB1FP	0-1	1.36	1	Yes

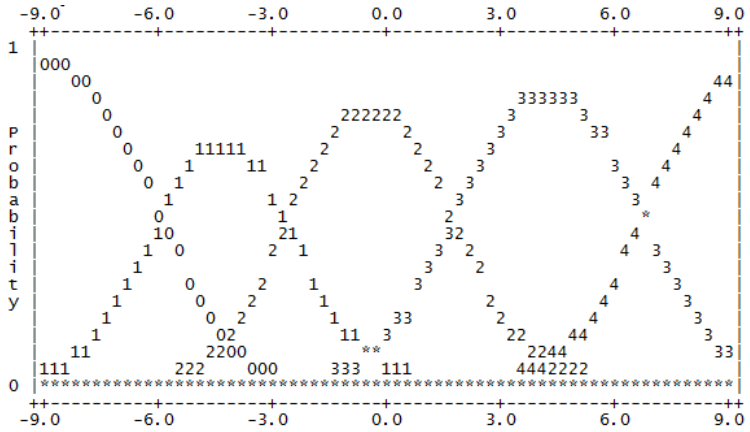
While the judge measurement report (Table 5.27.) tells us that the judges were not all the same in terms of harshness, we can see that they were consistent in their judgements. In fact judge Ou1 (indicating a person from outside the City & Guilds organization) was almost half a level lower than judge CG2 (from within the City & Guilds organisation or associated with the Communicator examination – e.g. a test task writer or examiner). In fact, with the exception of judge Ou1, all of the other judges were very much in agreement, with a maximum difference of only .23 (between CG1 and CG2).

Table 5.27. Judge Measurement Report (Writing)

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Nu	Rater
85	56	1.5	1.68	.29	.26	.9	0	.9	0	1	CG1
102	56	1.8	1.91	-.74	.27	1.2	0	1.1	0	2	CG2
90	56	1.6	1.75	.03	.26	.8	-1	.8	-1	3	CG3
78	56	1.4	1.45	1.09	.27	.7	-1	.7	-1	5	Ou1
91	56	1.6	1.77	-.10	.26	1.0	0	1.0	0	6	CG5
88	56	1.6	1.69	.25	.26	1.1	0	1.2	0	7	CG6
97	56	1.7	1.85	-.41	.26	.8	-1	.8	-1	8	Ou2
94	56	1.7	1.80	-.20	.26	.8	-1	.8	-1	9	CG7
94	56	1.7	1.80	-.20	.26	1.4	2	1.4	2	10	Ou3
91.0	56.0	1.6	1.74	.00	.26	1.0	-.3	1.0	-.4	Mean (Count: 9)	
6.6	.0	.1	.12	.49	.00	.2	1.4	.2	1.5	S.D.	

RMSE (Model) .26 Adj S.D. .41 Separation 1.57 Reliability .71
Fixed (all same) chi-square: 29.7 d.f.: 8 significance: .00

Figure 5.11. Probability Curves (Round 2 – Writing)



When we analysed the reliability of the judgements made we found that the Cronbach Alpha estimate was 0.976 and the inter-class correlation was 0.971. These figures support our claims that the writing standardisation was more than adequate from this perspective.

Commentary

The evidence from this part of the standard setting process suggests that there are clear links between the tasks indicated by the Council of Europe as being representative of the Level B2 and the tasks presented by City & Guilds for consideration by the judges in this event.

The data also suggest that these judges see a clear difference between levels of attainment within Level B2, with those task performances originally seen to be at the First Class Pass level being significantly higher than the other task performances.

Despite the lack of a larger sample of Level B2 task performances in writing, we feel that there is solid evidence from this standard setting event that the City & Guilds Communicator Writing paper is at the Level B2 and that the ratings of the performances reflect that level.

5.4. Claims

The Draft Manual (Council of Europe, 2003) suggests that at this point in the linking process, we should be in a position to make claims based on the specification and standardization phases. We have already stated (in Section 4.6.) that we feel uneasy about making claims based only on the completed specification forms, as they offer only a very

basic description of the test. We feel that any claims we make at this stage of the process (having completed the standardization stage) are only likely to be slightly stronger, but accept that since these claims are at least supported by empirical evidence they are more likely to be seen by the outside observer as being worthy of support.

Nevertheless, we would prefer not to make anything like a strong claim at this point in time, believing, as we do, that a more complete validation argument should be made to support any strong and meaningful claim of level.

We would therefore prefer to say simply that the evidence from the standardization process, that the Communicator examination's Reading, Listening and Writing papers are at CEFR Level B2 is strong enough for us to move to the final validation stage of the process.

This stage is presented in the following part of the report.

Part 6 – The Validation Stage

The Draft Manual (Council of Europe, 2003) appears to imply that the validation stage should be comprised of two sections, internal and external validity. This very limited view of the subject really is quite problematic in our view. Instead, we feel that the validation stage should be just that, a statement on the validity evidence which can be used to support any claim of a strong link between a given test and a specific CEFR level.

With this in mind, we turn to Weir's validation frameworks for an operational model of validation for the Communicator. The following elements of the framework are presented for each paper:

- The Test Taker
- Context Validity Evidence
- Scoring Validity Evidence
- Criterion Validity Evidence

6.1. The Test Taker

As mentioned earlier, City & Guilds, like many examination boards, routinely collects a range of information about its candidates. This information is used to monitor the population to ensure that there are no changes, such as some movement in the age or educational background of the candidature that might have an impact on the suitability of the test tasks or topics. The organisation also analyses the resulting data for evidence of differential item and test functioning. The variables for which data are routinely collected are:

- Linguistic background (L1)
- Language learning background
- Age
- Educational level
- Socio-economic background

- Social-cultural factors
- Ethnic background
- Gender

The topics and tasks in each of the three papers are designed with the test population in mind, see Appendix 5 for a full copy of a Communicator Paper. Care is taken to avoid topics that may negatively impact on candidates' performance while the reading and listening input texts are rigorously screened for content and language to ensure that they are appropriate for the typical candidature. The organisation is also keen to ensure that the cognitive load of the examination tasks is appropriate to the candidature. In order to ensure that this is the case, feedback is routinely gathered from centres and teachers and fed back into the test development system.

City & Guilds published their policy on special arrangements for students taking their examinations in 2007. For a broad understanding of how this policy impacts on individual candidates see that document (City 7 Guilds, 2007)

6.2. Context Validity Evidence

Context validity can be viewed from the perspectives of task settings (the conditions under which the task is performed), task demands (the linguistic demands of the input and expected output) and the administrative conditions (the non-language aspects of task administration). These are summarised in the following three tables.

Table 6.1. Overview of Context Validity – Task Settings (Reading)

Parameter	Description
Purpose	The requirements of the task. As with tests of other aspects of language ability this gives candidates an opportunity to choose the most appropriate strategies and determine what information they are to target in the text in comprehension activities. Facilitates goal setting and monitoring (key aspects of cognitive validity).
Response format	How candidates are expected to respond to the task (e.g. MCQ; SAF; Matching, handwriting, writing on computer etc.). Different formats can impact on performance.
Known criteria	As with listening tests, letting candidates know how their performance will be assessed. Means informing them about rating criteria beforehand (e.g. in SAF, is spelling or grammar relevant as is the case in IELTS; for writing, letting the test takers know about the assessment criteria before they attempt the task).
Weighting	Goal setting can be affected if candidates are informed of differential weighting of items before test performance begins. Items should only be weighted where there is compelling evidence that they are more difficult and/or more central to the domain.
Order of Items	In reading comprehension tests items will not appear in the same order as the

	information in the text where students search read (i.e. for scanning) but may appear in any order for careful reading..
Time constraints	Can relate either to pre-performance, or during performance. The latter is very important in the testing of reading, as without a time element we cannot test skills such as skimming and scanning (i.e. without this element all reading will be 'careful')

Table 6.2. Overview of Context Validity – Task Demands (Reading)

Parameter	**Description**
Discourse Mode	Includes the categories of genre, rhetorical task and patterns of exposition
Channel	In terms of input this can be written, visual (photo, artwork, etc), graphical (charts, tables, etc.) or aural (input from examiner, recorded medium, etc). Output depends on the ability being tested.
Text Length	Amount of input/output
Writer-reader relationship	This can be an actual or invented relationship. Test takers are likely to react differently to a text where the relative status of the writer is known – or may react in an unpredictable way where there is no attempt to identify a possible relationship (i.e. the test developer cannot predict who the test taker may have in mind as the writer and so the test developer looses a degree of control over the conditions)
Nature of Text(s)	This will include the rubric and tasks
Nature of Information	The degree of abstractness. Research suggests that more concrete topics/inputs are less difficult to respond to than more abstract ones.
Content Knowledge	Same as background knowledge which is very likely to impact on test task/item performance.
Linguistic	
Lexical Range	These relate to the language of the input (usually expected to be set at a level below that of the expected output) and to the language of the expected output. Described in terms of a curriculum document or a language framework such as the CEFR.
Structural Range	
Functional Range	

Table 6.3. Overview of Context Validity – Task Demands (Reading)

Parameter	**Description**
Physical Conditions	All of these elements are taken into consideration in the Information for Centres documents. Centres are routinely monitored to ensure that they are complying with the regulations.
Uniformity of Administration	
Security	

With these descriptions on mind we now go on to look at the context validity of the Communicator papers.

Table 6.1. Context Validity (Task Settings) of the Communicator Papers

OVERALL DESIGN	Listening	Reading	Writing
Purpose	General Proficiency		
Intended population	Late teens to early adult learners of English aiming for certification at CEFR B2		
	Data collected by C&G in CIS sheet (completed by each test taker)		
Intended decisions/Stakes	CEFR B2 ability claim, high stakes		
Response format	Combination of MCQ, SAF and matching	Combination of MCQ, SAF and matching	Handwritten response
Number of tasks	4	4	2
Task types	Task 1: 8 multiple choice items (1 per dialogue) each with 3 distractors. After listening to an incomplete dialogue candidates identify appropriate response. Task 2: 6 multiple choice items (2 per conversation) each with 3 distractors. After listening to the conversation candidates identify required answer to written prompt. Task 3: Note or message pad with headings, candidates listen to a monologue and select required information to complete notes. Task 4: Candidates listen to a dialogue and select a, b, c or d to answer questions or complete statements.	Task 1: A text with gaps in 5 sentences and a list of 8 items of text. Candidates insert the correct letter for an item of text in the relevant box. Task 2: 1 long text followed by 7 multiple choice comprehension items (includes 1 example). Candidates read and select correct response from multiple choice options by circling appropriate letter. Task 3: 4 short texts. Candidates use texts to find the correct answers to questions. Task 4: Candidates read text and answer questions to show understanding	Task 1: Formal report or article in response to written, graphic or visual input Task 2: A Letter, a narrative or a descriptive composition to produce a long continuous text on a single given topic.
Order of tasks	Order as in paper, test takers may respond in any order	Order as on recording	Order as in paper, test takers may respond in any order
Weighting of tasks	Equal weighting of tasks within each paper		
Rating Scale type	N/A (answer key)	N/A (answer key)	Task specific scales for both tasks (see Appendix 6 for an example)
Reporting type	As well as the global pass/fail/first class pass, students also receive the following information about their subtest performance: • Grade or p/f/fcp per subtest Profile of aspects of performance per subtest using a Performance Code Report		
Assumptions re population			
Background Knowledge	Broad candidature so this is dealt with by selecting only clear topics accessible to the general reader.		
Language Knowledge	Candidates are expected to be at the CEFR B2 level (so input will be at B2 or below)		

Table 6.2. Context Validity (Task Demands) of the Communicator Paper

Parameter	Listening	Reading	Writing
Discourse Mode	Task 1: Adult conversations, formal & informal Task 2: Conversation Task 3:Phone call / Broadcast / lecture Task 4: conversation	Task 1: Discursive, explanatory, descriptive or biographical text. Task 2: Expository, news story, article, report, review or proposal. Task 3: short text, common theme, different text types, including diary entry, notice, email, news story, letter, article, memo, proposal etc. Task 4: A narrative, discursive, explanatory, descriptive, biographical, or instructive text.	Task 1: Formal report or article Task 2: Letter
Channel	Aural	Written	Written
Text Length	Task 1: Maximum of 25 words per dialogue input Maximum of 30 words for each set of multiple choice options. Task 2: 100 – 160 words per conversation. Maximum 12 words per stem, maximum 30 words for each item Task 3: 440-480 words spoken text. Maximum 3 words for each answer. Task 4: 540-600 words for dialogue, maximum turn length 50 words. 12 words maximum per stem, maximum 30 words for each set of multiple choice options	Task 1: 260 – 300 words in text including answers A – H. Task 2: Text length 400 – 420 words including title. Multiple choice items: each stem 1 – 8 words, each option 3 – 15 words. Task 3: 80 – 90 words per text. Items 1 – 5: 2 to 12 words each. Items 6 – 10: 5 – 15 words each. Task 4: 380 – 450 words in text including title. Maximum 10 words in each stem.	Task 1: 50 – 65 words in prompt text, candidate to write 100 – 150 words Task 2: 100 – 150 words to be written by candidate.
Writer-reader relationship	Task 1: Unspecified relationship Task 2: Unspecified relationship Task 3: Unspecified relationship Task 4: Unspecified relationship	Task 1: Unspecified writer Task 2: Specified: Known (high Status – e.g. report from manager, article from academic etc). Task 3: Mix of known and unknown writers; mix of specified and unspecified writers. Task 4: Unspecified writer.	Task 1: Specified, a mixture of known and unknown. Task 2: Specified & known audience
Nature of Text(s)	Task 1: Short conversations on concrete and abstract topics Task 2: Conversations on concrete and abstract topics Task 3: Monologue such as lecture, broadcast, presentation etc. Task 4: Discussion on concrete and abstract topics	Task 1: Concrete or Abstract Task 2: Concrete or Abstract Task 3: Wide range of topics – general, social, work, study. All 4 texts should be linked by a common theme. Task 4: concrete or Abstract	Task 1: Formal report or article Task 2: Letter
Nature of Information	Concrete only		
Lexical Range	All of these have been developed based on the descriptors in the CEFR for Level B2.		
Structural Range	In addition there are extensive lists of syntax for all C&G ESOL examinations		
Functional Range	Research also into the lexical profile of reading texts (Schmitt, 2007) used in developing input texts and all items		

6.3. Scoring Validity Evidence

The key areas of scoring validity for reading and listening are:

- Accuracy of the answer key
- Item performance
- Internal Consistency
- Standard Error of Measurement
- Marker Reliability

In order to collect the data to respond to these parameters, City & Guilds undertook to operationalise the newly changed test with a population of approximately 330 candidates who were deemed by their school and/or test centre to be ready to sit the Communicator examination. This work was undertaken in late 2007 and the results used to feed into this aspect of the project. At the same time, candidates were asked to respond to additional tasks (in both reading and listening) that were suggested by the Council of Europe to be at this level.

Table 6.3. Scoring Validity of the Communicator Paper (Listening and Reading)

Parameter	Listening	Reading
Accuracy of the answer key	Systematically checked on production of task, then again both pre and post test administration	
Item performance	Ave. Item Facility = 50.86 Ave Item Disc. = 0.36	Ave. Item Facility = 48.84 Ave Item Disc. = 0.36
Internal Consistency	0.81 (N=330)	0.77 (N=330)
Standard Error of Measurement	1.73 Candidates within 2 points of the cut score will automatically have their scores reviewed	1.94 Candidates within 2 points of the cut score will automatically have their scores reviewed
Marker Reliability	Optical Mark Reader (OMR) is used to capture test scores – expected reliability is 99.98%	

Table 6.4. Scoring Validity of the Communicator Paper (Writing)

Parameter	Writing
Rating Scale	As mentioned earlier in this report, the rating scale used for Communicator was developed based directly on the descriptors at Level B2 of the CEFR. The indications from the trials are that the scale is working well (based on a multi-faceted Rasch analysis of rater data.
Rater Selection	Minimum requirements for rater selection are set out in the Communicator guidelines. These refer to teaching and where possible testing experience at level B2.
Rater Training	All raters are routinely trained using materials based on the CEFR and now using Council of Europe Recommended Task Performances.
Rater Monitoring	Raters are routinely monitored during the year to ensure they are on level. In addition City & Guilds regularly measures both rater agreement and intra-rater reliability. The most up-to-date data follows.
Rater Agreement	Inter-class correlation for the writing was estimated at 0.971
Rater Consistency	No raters appeared inconsistent in our study (based on MFR infit mean square statistics being in the acceptable range of 0.5 to 1.5).
Estimated SEM	0.181
Rating Conditions	Raters may mark scripts in their own work environment, though they are given clear and strict instructions relating to the conduct of the assessment (e.g. guidelines for best practice).
Grading and Awarding	Since Communicator is an 'on-demand' examination all care is taken to ensure the reliability and fairness of the scoring system. Where there are issues with a score, these are taken up with City & Guilds for monitoring.

6.4. Criterion-Related Validity Evidence

In order to gather criterion-related evidence of validity, a study was commissioned in which a large population of test candidates were asked to sit a reading and listening paper and also a number of Council of Europe recommended tasks at CEFR Level B2. Candidates were also asked to complete 'Can-Do' questionnaires for both reading and listening.

The reading element of this study is reported in the Section 6.4.1. and the listening element in Section 6.4.2.. The evidence for the writing paper is based on the performances rated during the standard setting study described in the previous part of this report, and is reported separately in Section 6.4.3.

Before reporting on the findings of the study, we should briefly describe the population for the reading and listening papers.

Population

The population for the reading and listening papers consisted of 397 candidates at 17 centres across Europe. All candidates were asked to take a reading and listening paper and complete 'can-do' questionnaires as described above. Two versions of the reading paper were used in this study and, since the same Council of Europe recommended tasks for reading were used by all candidates, it was hoped that these would be used to link the two papers for the final analysis reported. However, there were too few items in the recommended task to allow us to anchor the two test versions, so the following section reports only on the first, and larger of the two tests. This was chosen both because the size of the population was larger and because these candidates took four City & Guilds tasks and two recommended tasks.

When the final dataset was analysed it was found that we had a complete set of responses for 200 candidates, so this population was chosen as the final set for the project.

All candidates were attending language schools or centres where the Communicator was a regular final target examination. The candidates who sat these papers were all seen by their centres to be at the level of Communicator and were identified by their teachers as being ready to sit the test. They were also seen as representative of the usual Communicator candidature in terms of gender spread, age and educational experience.

6.4.1. The Criterion Study: The Reading Paper

The participants were asked to sit for four City & Guilds Communicator tasks and two recommended tasks. The reason for the reduced length of the Communicator test was to lessen the load on the participants. The data were first analysed using classical statistics and later using IRT.

The descriptive statistics (Table 6.5. confirm that there appears to be little meaningful difference between the average score on the City & Guilds tasks and the scores achieved for the other tasks. The tasks used for this study came from Cambridge ESOL (CambR) and from the Finish Ministry of Education (FinnR). Note that all mean scores are presented on a continuum of 0 to 1, with the latter indicating a perfect score.

Table 6.5. Descriptive Statistics for Reading

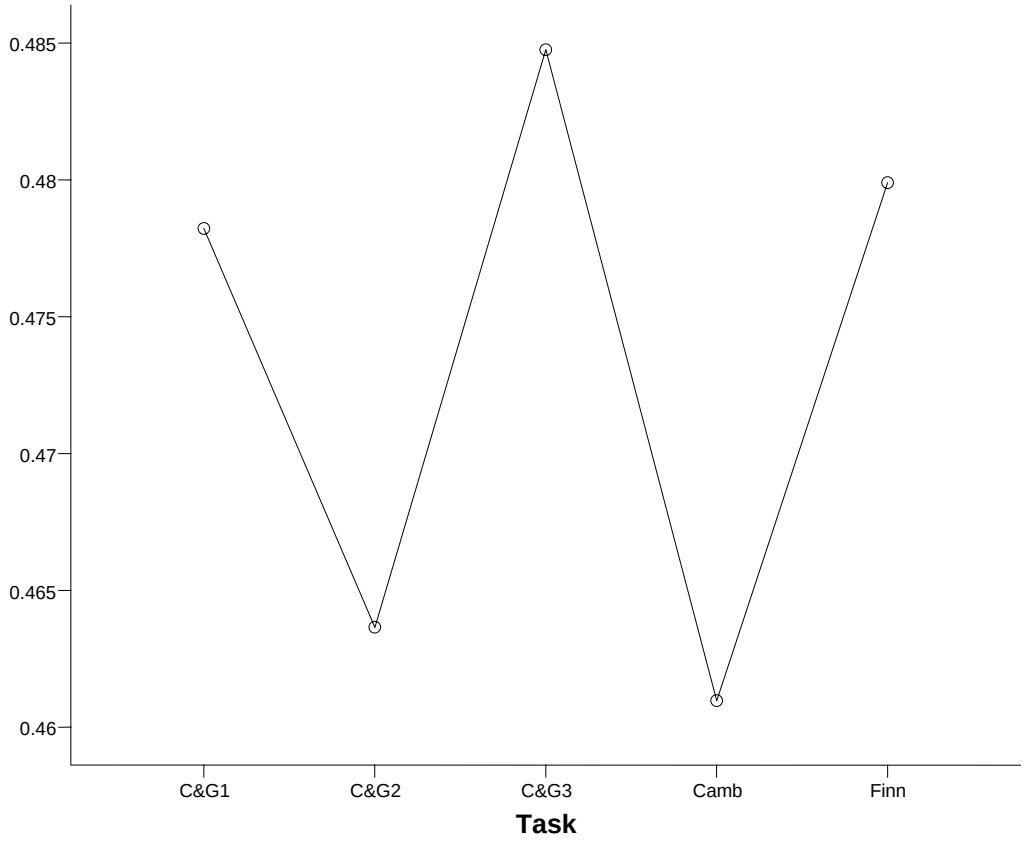
Descriptives

Average Score

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
C&G1	199	.4782	.19556	.01386	.4509	.5056	.17	.83
C&G2	199	.4637	.21481	.01523	.4336	.4937	.17	.83
C&G3	199	.4848	.20953	.01485	.4555	.5140	.17	.83
Camb	199	.4610	.26734	.01895	.4236	.4983	.17	1.00
Finn	199	.4799	.18862	.01337	.4535	.5063	.25	1.00
Total	995	.4735	.21672	.00687	.4600	.4870	.17	1.00

These numbers suggest that there is very little meaningful difference between the City & Guilds Tasks and the recommended tasks. The chart (Figure 6.1.) also suggests that there are little differences between the tasks – while the chart appears to show a very jagged profile of performance, the scale to the left (which indicates mean performance on a scale of 0 to 1) shows that the differences are, in fact very small.

Figure 6.1. Chart of Mean Performance on Reading Tasks



The one-way ANOVA (Table 6.6.) supports this assertion and indicates that there is no significant difference between the five test tasks. Post hoc analysis supports this finding, see Table 6.7.

Table 6.6. One-Way ANOVA for Reading

ANOVA

Average Score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.088	4	.022	.469	.758
Within Groups	46.597	990	.047		
Total	46.685	994			

Table 6.7. Post Hoc (Scheffe) from ANOVA for Reading

Multiple Comparisons

Dependent Variable: Average Score

Scheffe

(I) Task	(J) Task	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
C&G1	C&G2	.01457	.02175	.978	-.0525	.0817
	C&G3	-.00653	.02175	.999	-.0737	.0606
	Camb	.01725	.02175	.960	-.0499	.0844
	Finn	-.00168	.02175	1.000	-.0688	.0654
C&G2	C&G1	-.01457	.02175	.978	-.0817	.0525
	C&G3	-.02111	.02175	.918	-.0882	.0460
	Camb	.00268	.02175	1.000	-.0644	.0698
	Finn	-.01625	.02175	.968	-.0834	.0509
C&G3	C&G1	.00653	.02175	.999	-.0606	.0737
	C&G2	.02111	.02175	.918	-.0460	.0882
	Camb	.02379	.02175	.879	-.0433	.0909
	Finn	.00486	.02175	1.000	-.0623	.0720
Camb	C&G1	-.01725	.02175	.960	-.0844	.0499
	C&G2	-.00268	.02175	1.000	-.0698	.0644
	C&G3	-.02379	.02175	.879	-.0909	.0433
	Finn	-.01893	.02175	.944	-.0860	.0482
Finn	C&G1	.00168	.02175	1.000	-.0654	.0688
	C&G2	.01625	.02175	.968	-.0509	.0834
	C&G3	-.00486	.02175	1.000	-.0720	.0623
	Camb	.01893	.02175	.944	-.0482	.0860

The evidence from this study indicates that the three City & Guilds Communicator tasks performed by these candidates are generating similar in levels of performance to the two tasks recommended by the Council of Europe. We can therefore say that the three City & Guilds Communicator tasks undertaken here are as likely to be at CEFR Level B2 as those tasks recommended by the Council of Europe.

6.4.2. The Criterion Study: The Listening Paper

Unlike the reading paper, there was a larger population for the listening. This time there were enough items in the Finnish task to successfully link the two sets of data gathered for the project so the overall population was 330.

The descriptive statistics for these data show that the situation with the listening paper is quite different to that of the reading. Here (Table 6.8.) we can see that there are a number of differences between the tasks.

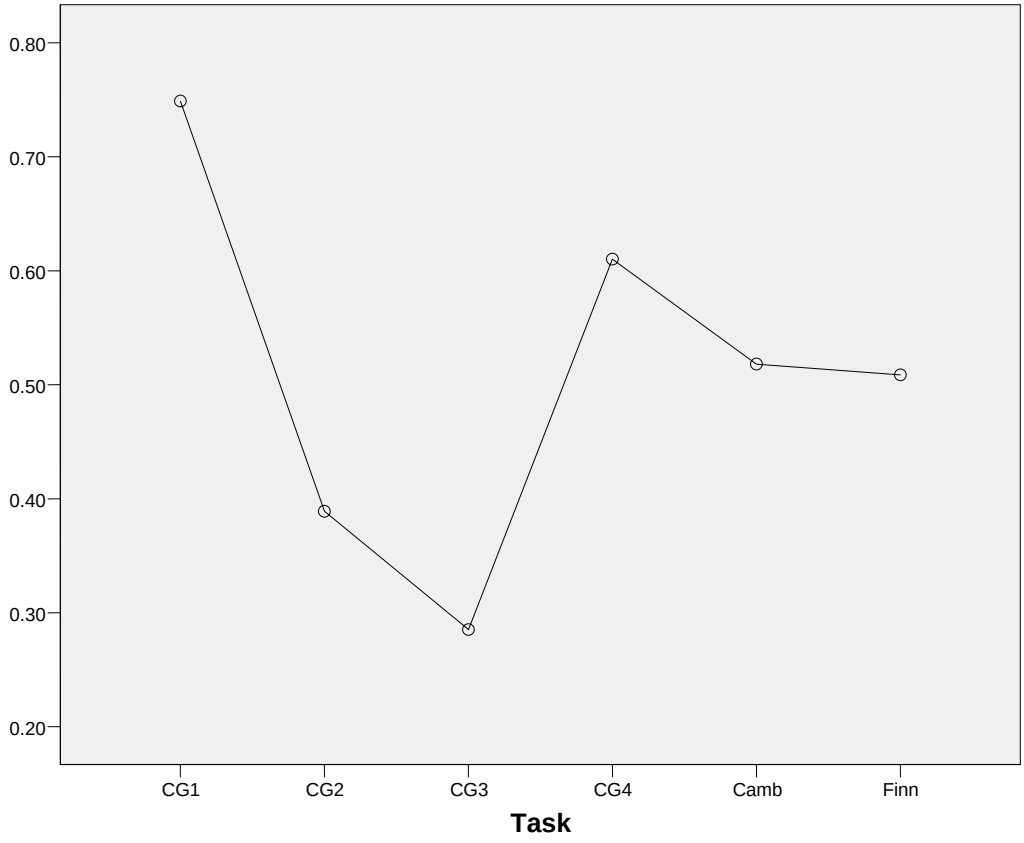
Table 6.8. Descriptive Statistics for Listening (All C&G Tasks)

Descriptives

Score

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
CG1	330	.7489	.19382	.01067	.7279	.7699	.00	1.00
CG2	330	.3890	.21867	.01204	.3653	.4127	.00	.88
CG3	330	.2852	.22635	.01246	.2607	.3097	.00	.88
CG4	330	.6102	.26622	.01465	.5813	.6390	.00	1.00
Camb	330	.5182	.33802	.01861	.4816	.5548	.00	1.00
Finn	330	.5087	.23221	.01278	.4835	.5338	.00	1.00
Total	1980	.5100	.29072	.00653	.4972	.5228	.00	1.00

Figure 6.2. Chart of Mean Performance on Listening Tasks (All C&G Tasks)



The one-way ANOVA results table (Table 6.9.) confirms that there are significant differences between the tasks, and the follow-up post hoc analyses confirms that there are a range of differences to be found.

Table 6.9. One-Way ANOVA for Listening (All C&G Tasks)

ANOVA

Score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	43.665	5	8.733	139.480	.000
Within Groups	123.596	1974	.063		
Total	167.261	1979			

Table 6.10. Post Hoc (Scheffe) from ANOVA for Listening (All C&G Tasks)

Multiple Comparisons

Dependent Variable: Score

Scheffe

(I) Task	(J) Task	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
CG1	CG2	.35985*	.01948	.000	.2950	.4247
	CG3	.46364*	.01948	.000	.3988	.5285
	CG4	.13869*	.01948	.000	.0738	.2036
	Camb	.23068*	.01948	.000	.1658	.2956
	Finn	.24021*	.01948	.000	.1753	.3051
CG2	CG1	-.35985*	.01948	.000	-.4247	-.2950
	CG3	.10379*	.01948	.000	.0389	.1687
	CG4	-.22116*	.01948	.000	-.2860	-.1563
	Camb	-.12917*	.01948	.000	-.1940	-.0643
	Finn	-.11964*	.01948	.000	-.1845	-.0548
CG3	CG1	-.46364*	.01948	.000	-.5285	-.3988
	CG2	-.10379*	.01948	.000	-.1687	-.0389
	CG4	-.32495*	.01948	.000	-.3898	-.2601
	Camb	-.23295*	.01948	.000	-.2978	-.1681
	Finn	-.22343*	.01948	.000	-.2883	-.1586
CG4	CG1	-.13869*	.01948	.000	-.2036	-.0738
	CG2	.22116*	.01948	.000	.1563	.2860
	CG3	.32495*	.01948	.000	.2601	.3898
	Camb	.09199*	.01948	.000	.0271	.1569
	Finn	.10152*	.01948	.000	.0366	.1664
Camb	CG1	-.23068*	.01948	.000	-.2956	-.1658
	CG2	.12917*	.01948	.000	.0643	.1940
	CG3	.23295*	.01948	.000	.1681	.2978
	CG4	-.09199*	.01948	.000	-.1569	-.0271
	Finn	.00952	.01948	.999	-.0554	.0744
Finn	CG1	-.24021*	.01948	.000	-.3051	-.1753
	CG2	.11964*	.01948	.000	.0548	.1845
	CG3	.22343*	.01948	.000	.1586	.2883
	CG4	-.10152*	.01948	.000	-.1664	-.0366
	Camb	-.00952	.01948	.999	-.0744	.0554

*. The mean difference is significant at the .05 level.

In order to further investigate the situation, we decided to average the four City & Guilds tasks to see if the combination is more indicative of the level as defined by the recommended tasks. The descriptive statistics (Table 6.11.) suggest that this is indeed the case, a situation also apparent in the chart (Figure 6.3.) with the average Cambridge ESOL task being very slightly easier than the average scores for either the City & Guilds tasks or the Finnish tasks (though it should be noted that any differences here are not statistically significant).

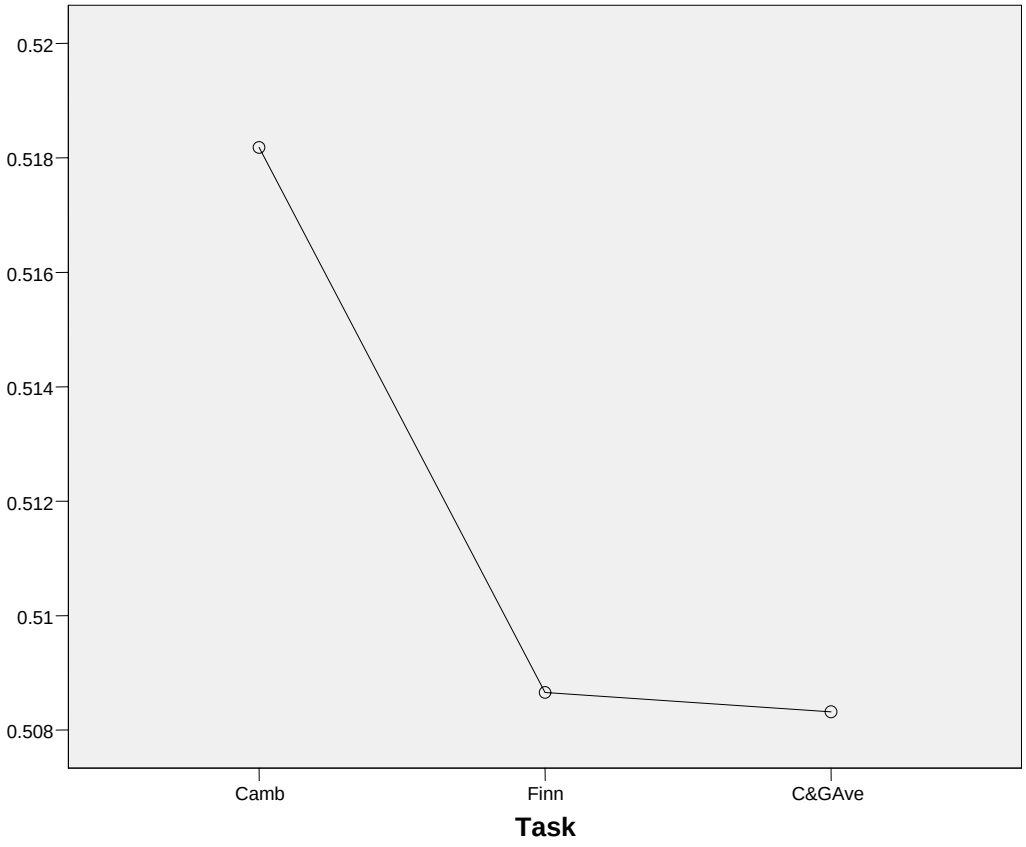
Table 6.11. Descriptive Statistics for Listening (Using Average C&G Score)

Descriptives

Score

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Camb	330	.5182	.33802	.01861	.4816	.5548	.00	1.00
Finn	330	.5087	.23221	.01278	.4835	.5338	.00	1.00
C&GAve	330	.5083	.16718	.00920	.4902	.5264	.13	.91
Total	990	.5117	.25547	.00812	.4958	.5277	.00	1.00

Figure 6.3. Chart of Mean Performance on Listening Tasks (Using Average C&G Score)



The one-way ANOVA results (Table 6.12) indicate that the differences observed above are indeed, not statistically significant. This finding is re-iterated in the post hoc results which show clearly that there are no significant differences between the average performance on the three sets of tasks.

Table 6.12. One-Way ANOVA for Listening (Using Average C&G Score)

ANOVA

Score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.021	2	.010	.158	.854
Within Groups	64.527	987	.065		
Total	64.548	989			

Table 6.13. Post Hoc (Scheffe) from ANOVA for Listening (Using Average C&G Score)

Multiple Comparisons

Dependent Variable: Score

Scheffe

(I) Task	(J) Task	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Camb	Finn	.00952	.01991	.892	-.0393	.0583
	C&GAve	.00986	.01991	.885	-.0389	.0587
Finn	Camb	-.00952	.01991	.892	-.0583	.0393
	C&GAve	.00034	.01991	1.000	-.0485	.0491
C&GAve	Camb	-.00986	.01991	.885	-.0587	.0389
	Finn	-.00034	.01991	1.000	-.0491	.0485

Like with the reading paper, we can therefore argue that the overall level of ability represented by successful completion of the City & Guilds tasks used in this project is similar to that represented by the recommended tasks from both Cambridge ESOL and the Finnish Ministry of Education. This suggests that the overall CEFR level of the Communicator listening paper is as likely to be at B2 as the recommended tasks.

However, it is clear that the different tasks within the Communicator represent a wide range of listening ability. This finding reflects the findings of the Finnish Ministry of Education in the documentation that they supplied with their recommended listening tasks.

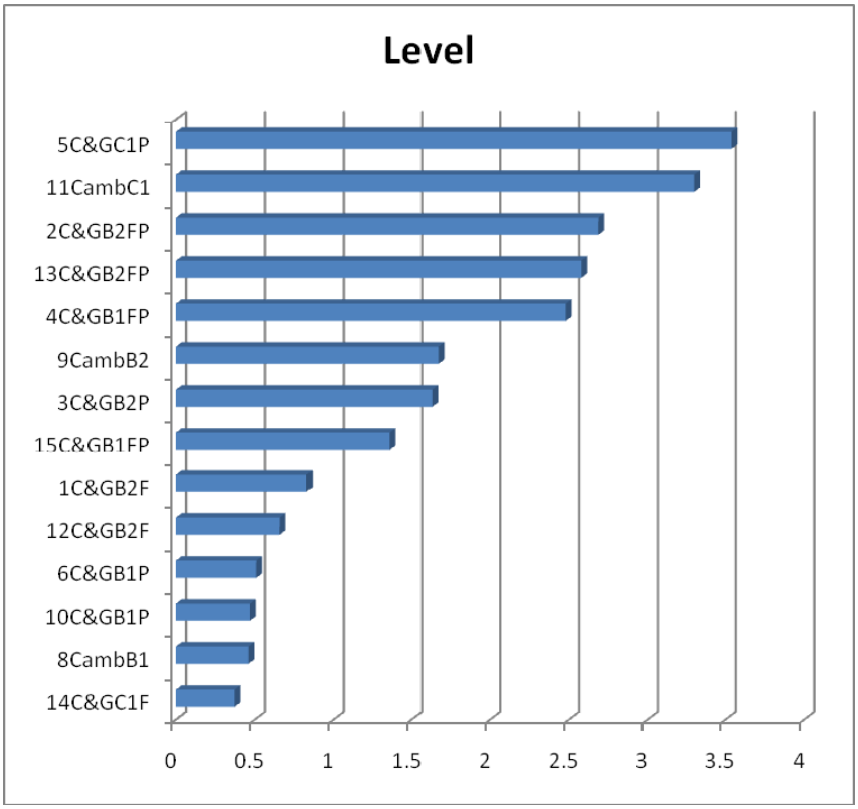
6.4.3. The Criterion Study: The Writing Paper

Unlike the case with the reading and listening papers, there were no additional performances collected from Communicator candidates for the writing criterion study. Instead, we decided to re-visit the data collected as part of the standard setting event, in which a number of experienced test developers and City & Guilds test writers and examiners rated a range of test task performances. These tasks were focused mainly on the CEFR Level B2 though we included both B1 and C1 task performances also.

Criterion-Related Evidence

If we look back to Table 5.25. and the summary chart for the writing MFR analysis (Figure 5.9.) we can see clear evidence that the City & Guilds tasks are very much on a level with the single set of recommended tasks from Cambridge ESOL.

Figure 6.4. Chart of Judgements of Writing Tasks



Looking again at the tasks (Figure 6.4.) we see that they appear to fall into the following categories:

C1 – 5C&GC1P & 11CambC1

B2.2 – 2C&GB2FP, 13C&GB2FP & 4C&GB1FP

B2.1 – 9CambB2, 3C&GB2P, 15C&GB1FP

B1.2 – 1C&GB2F & 12C&GB2F

B1.1– 6C&GB1P, 10C&GB1P, 8CambB1 & 14C&GC1F

We believe that this should be seen as strong evidence that the Communicator tasks at CEFR Level B2 are clearly at that level (as represented by the single recommended task available in the public domain at the time of this project). However, we would urge some caution as the level is only defined by a single recommended task at this point in time. Until we have more and better defined task performances at each CEFR level we cannot hope to accurately model the levels. In addition, it is important that we should ask candidates to perform both the focus tasks (i.e. the tasks that represent the test we are attempting to benchmark to the CEFR) and a range of tasks that are representative of the level.

6.5. Claims

In terms of the original draft manual, it should be possible to make claims of different sorts at the end of each stage of the linking process. However, we have argued in this report that at these stages the amount and nature of evidence make these claims weak at best.

Now that we have completed all four stages of the process (though it should be recognised that the Familiarisation stage is not really an independent stage at all but contributes to all of the other stages), we feel we are in a position to make a strong claim of a link between the City & Guilds Communicator examination and the level of language described at B2 of the CEFR and reflected in the various tasks recommended by the Council of Europe as representing this level.

The evidence to support this was presented in the form of:

- Successful completion of the Specification Forms (A1-A23), which demonstrated the technical quality of the examination, and the quality of the system that supports it.
- Successful standard setting events for the Listening and Reading papers, which indicated that the cut scores could be directly linked to the definition of the minimally competent candidate at CEFR Level B2 in both skill areas. In addition,

the judgements of the expert panel indicated a clear link between the Communicator (and other City & Guilds tasks) and the recommended tasks, and also evidence that all are clearly at the level expected. It should again be remembered that the main focus here was the B2 level tasks, so that any evidence of a link at C1 or B1 should be viewed with caution.

- Successful presentation of a range of evidence of the validity of the Communicator papers. This evidence was not presented as relating to either *internal* or *external* validity, as suggested it the manual. This is because we do not see this approach to validation as being particularly helpful or informative. Instead, we have used the Weir (2005) frameworks, which meant that we were able to present the type of evidence called for in the manual but in a way that formed part of a broader validation argument. We believe that this method of presenting the validity evidence is far more informative, and links together the initial specification phase to the validation phase. This is particularly the case as the specifications for the Communicator (and all other City & Guilds ESOL and ISESOL examinations) use the Weir frameworks as a model for specifications.

Part 7 – Summary and Discussion

In this part of the report we will attempt to bring together the various strands of the process and to reflect on what it has meant both for the Communicator examination itself and for City & Guilds as an institution.

7.1. Summary of the Project

In this project we set out to establish empirical evidence of a link between the Communicator examination and CEFR Level B2. The test itself had been specifically designed with this level in mind, and at all stages of the development process developers, syllabus writers, task & item writers and administrators were familiarised with the CEFR and were constantly instructed to refer to it.

The methodology of the project was dictated by the Pilot Manual (Council of Europe, 2003), though it became clear as the project proceeded that the process outlined in that document was not likely to be without problems.

The stages of the project suggested in the Pilot Manual were

Familiarisation	with the CEFR
Specification	of the test using a set of forms
Standardisation	setting cut scores and establishing a link between these and the CEFR level
Validation	internal (psychometric qualities) and external (criterion-related evidence)

It became apparent to the project team that there were at least two key issues that had not been addressed explicitly in the Pilot Manual, these were:

1. The test itself should be critically reviewed and only then should any linking be undertaken.
2. Unless the processes of benchmarking are embedded in the systems of the institution, the whole reason for undertaking such a project is undermined. In other words, the exam should present a stable standard over the years, so there

has to be a mechanism to maintain the standards. This mechanism should have clear links to the CEFR level at which the examination is aimed.

For these reasons we decided to broaden the project to include additional elements. These were the inclusion of an additional review panel before the process began, and a widening of the familiarisation training to all members (including marketing and administration) of the City & Guilds ESOL and IESOL teams.

The project began in late 2006, with the systematic familiarisation of the City & Guilds staff. As this progressed, the initial expert panel meeting was held. At this point, the focus was on a detailed critical analysis of the Communicator, looking at the quality of the test as a whole, as well as undertaking detailed qualitative item and task analysis, and also looking at the probable level of these tasks and items.

Following this review, it was decided to undertake the following:

- Slightly update two of the reading tasks to better reflect the level
- Make some minor changes to the presentation of another of the reading tasks (removing formatting which it was felt impacted on the difficulty of the task)
- Slightly update two of the listening tasks (one of these was a change to the number of times candidates would listen to an input text)
- Review the training and standardisation of the writing raters.
- Update item writer guides to more forcefully reflect the level.
- Completely rewrite the test specifications so that they would be more easily interpretable by developers, and so that aspects of validation would be introduced at this early stage of the process.

All of this work was completed by spring 2007, with new task versions trialled and their quality and level established and the new specifications completed, though it should be said that these were working versions which were only finally completed in mid 2008.

At this point work began on the completion of the Specification forms. This proved time consuming and not always as informative as might have been hoped. The rationale behind having these forms is certainly sound, with sections requiring some level of detail about the content and focus of the test papers. However, as they are not linked to any specific validation model, there appears to be no clear theoretical justification behind them.

When we were satisfied that the stages to date had been successfully completed, we moved on to the Standardisation stage. Tasks and performances were gathered to reflect the latest changes (though minor) to the test and formal standard setting events organised in Autumn 2007 and Spring 2008. These events included expert panel members from within City & Guilds and from outside the organisation and were charged with establishing evidence of a link between an agreed minimally competent candidate at CEFR Level B2 and the cut scores for the Communicator reading and listening papers. They were also asked to make judgements on the link between test task performances that had been recommended by the Council of Europe as being representative of the B2 Level while at the same time indicating that they felt the tasks and associated performances were indeed at that level. The success of these events meant that we could move forward to the Validation stage. The data for this stage were collected even as the previous stage was progressing.

The presentation of the validation evidence was undertaken using Weir's (2005) validation frameworks as a theoretical basis, which allowed us to systematically detail all aspects of the test system that contribute to its validity. This differs in the approach suggested in the Manual, which appears to focus on the psychometric qualities of the test only – an aspect of what Weir sees as Scoring Validity.

7.2. Summary of the Main Findings and Claims

The main findings of the project can be summarised as follows:

1. It was found that in order to claim a link to the CEFR at Level B2 the cut score for a passing grade for the Communicator Reading paper should be set at 15 (from a maximum of 30). The same cut score was recommended for the Communicator Listening paper. This is actually in line with current practice for Communicator.

2. Passing levels for the Communicator Writing paper were found to be in line with the Council of Europe recommended tasks for CEFR Level B2. The recommendation is that the cut level for this decision should not be altered at this point in time.

3. The linking process is long and demanding, both at the individual and institutional level. The complexity of the design means that it is expensive for any

institution to undertake, certainly to the extent undertaken by City & Guilds in this project. While this perhaps explains the reluctance of many examination boards to undertake a full linking project, we nevertheless recommend that the process be extended to as many of the other examinations in the ESOL suite as feasible.

4. Unless the test which is the focus of the linking project is shown to be robust in terms of quality and level, there is no point in even starting a linking project, as the process is unlikely to succeed beyond the standardisation stage without serious issues emerging. In fact, we feel that with a more demanding specification phase, issues should emerge more clearly at this early stage.
5. Limiting the validation evidence to estimates of internal and external validity is far too simplistic a view of validation. The CEFR should be demonstrated to impact on all aspects of the test, from the test taker to the task to the psychometric qualities and relative meaning or value of the test score.

Based on this project, it is the belief of the project team that the evidence presented here supports the claim that the Communicator tests English ability at CEFR Level B2.

7.3. Implications of the Project

This project has a number of important implications, for the focus examination itself, for the ESOL Suite of which the Communicator is just one of six examinations, for the institution and of course for the Pilot Manual.

7.3.1. For the Communicator Examination

We feel that the process of linking the Communicator examination to the CEFR, has resulted in systematic and sustainable improvements to the test and to the system that supports the test.

It is clear to us that the process has resulted in a test that is more clearly at level, is sound from an internal psychometric perspective and is more replicable and of a high quality. However, that is not all. The systems that support the examination have also been systematically improved and more explicitly linked to the CEFR. The item writers' guidelines are, we believe, up-to-date and more robust than in the past. The

specifications are now more likely to result in accurate replication of tests on level – one criticism of the old specification was the lack of detail and exemplification, this appears to have led to a tendency to drift away from the level. This is a warning for other test developers, who take time to specify their tests but do not routinely review these specifications (and their use) to ensure that there is no level or construct drift.

We now feel that we are in a position to consider suggesting a number of Communicator tasks to the Council of Europe for use as recommended level indicators in future linking projects.

7.3.2. For City & Guilds and Other Examination Boards

This is just the beginning of the process. A decision has been made to investigate the extension of this project to the other examinations in the ESOL Suite. Preliminary work has already begun to establish review panels so that the quality of the Communicator can be replicated across the Suite.

As we have seen above, the processes that were developed during this project have been embedded in the City & Guilds test development system, with the link to the CEFR for each examination linked to all aspects of the development and validation process. We believe that this approach is vital to ensure the sustainability of the examinations they develop in terms of validity in general and level in particular.

7.3.3. For the CEFR Linking Manual

The implications for the linking manual relate to the model or framework suggested in the manual, and reproduced here in Part 2, Figure 2.1. and to the tasks and performances that are claimed to reflect the different levels of the CEFR.

The Linking Process Model

In that model, the implication is that the process is essentially linear in nature. This means that the expectation of a linking project is that we should start with familiarisation, then move on to specification, standardisation and finally validation. It was clear from the beginning of this project, however, that an additional element should be added to the model. It seems to be expected in the original model that the test we are attempting to link is stable, on level and of a high quality. However, this situation is rarely the case in

reality, with many even well-known examinations not subjected to the kind of rigid scrutiny (in terms of the three criteria) we would expect.

We therefore suggest that any test that is to be the subject of a linking project should first undergo a systematic and critical review, ideally from a panel of experts in the area. This expert panel should be comprised of participants from both within the developing institution, with experience in the development and/or operational administration of the test in question, and outside of the institution. It is essential that all panel members be very familiar with the CEFR at the level of the test being linked and of the levels above and below that level.

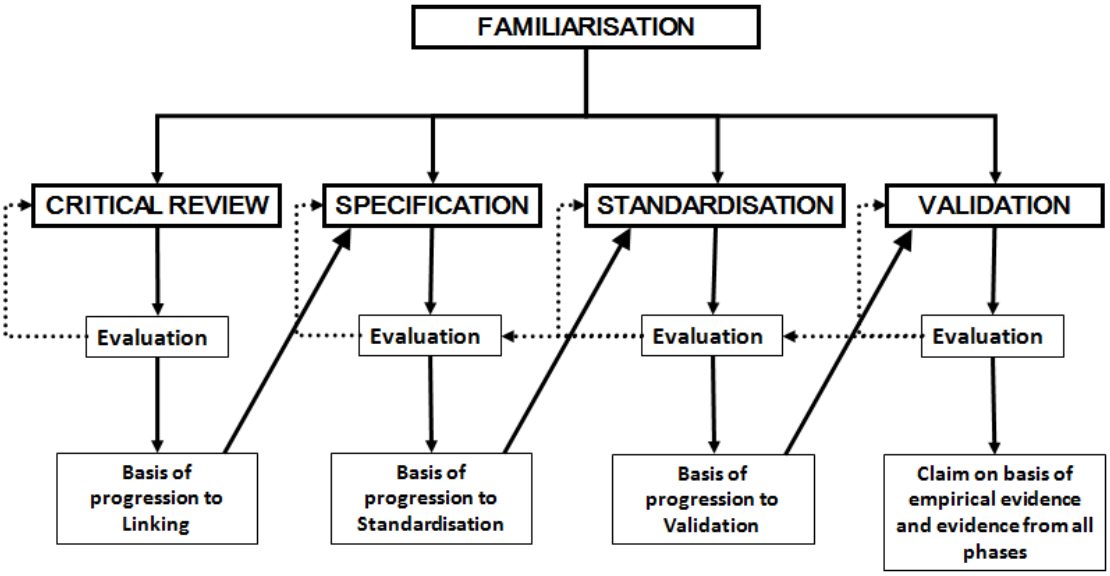
The advantage to doing this means that any issues with the test can be addressed and the systems that support the test (e.g. guidelines for item writers, regulations for administrators etc.) can be fully tested and, where needed, updated. This means that the move to the specification stage of the linking process will be smooth and the completion of the forms will be made much easier. In addition, the standardisation stage will have some hope of passing off without any major problems. It became clear to us during this project that it is at this standardisation stage that issues with test quality and level will become obvious.

It also became clear to us that the existing model needs to be updated to reflect the iterative nature of the process. We found that we were constantly evaluating each stage and considering how the findings might affect the work already done on the project – which in practice meant that we were returning to the earlier stages regularly during the process. For example, in terms of this particular project, we found in the critical review that there were some issues with the existing test system (in terms of the quality and level of a small number of items and also in terms of the support systems). If we had not conducted the review and gone ahead with the specification stage and then moved to the standardisation stage, these issues would have emerged only at this stage. This means that we would have to return to the specification stage, to reflect any changes (no matter how minor) to the test. This highlights a weakness of the specification stage (which requires no empirical evidence of level and asks a series of questions which require only a descriptive response – and is open to either intentional or unintentional abuse) and of the whole model.

We therefore suggest that the original model for linking be updated to reflect our experiences. The model we suggest can be seen in Figure 7.1. below. Here, we see that

there is an additional stage to the process, in which a critical review of the examination is first carried out in order to ensure that the test is working well and has the attributes (e.g. reliable and valid) that will make any linking meaningful – linking any test to the CEFR that is not of a sufficiently high quality will result in a meaningless claim. This also implies that we need to demonstrate the quality of a test along with any formal claim of a link.

Figure 7.1. Alternative Model for Linking a Test to the CEFR



We also feel that the notion of continuous evaluation of progress throughout the linking process should be stressed. Even when the critical review has been completed and any changes that are recommended from this review are in place, we should evaluate the process and the product of the review. If it is found that an item or task type needs to be updated, the changes should be evaluated empirically – as was done in the case of the Communicator reading and listening tasks. If we then feel that there is enough evidence that the quality and level of the test are likely to be acceptable, we would only then move to the specification stage.

Like the review stage, we suggest that the specification forms should be critically reviewed on completion and only when the linking institution is satisfied that the specification forms are an accurate reflection of the test and its supporting systems should the decision be made to move on to the standardisation stage. If there are issues found at this stage (for example, if some responses are felt to be weak or unclear this

may reflect some problem with the test itself) we should return to the review stage to identify exactly what the problem or problems might be.

Having then addressed these problems, the solutions should be evaluated and only then should a move to the next stage be considered. This constant reflection and evaluation means that the process is far from linear. Instead we should be constantly considering how the findings of the project impact on the test and visa versa. Therefore, by the time we reach the end of the process, we can be quite certain that the test we have attempted to link to the CEFR, is not only linked, but the link is meaningful because the test is of a high enough quality.

We also feel strongly that the limiting of validation evidence to internal and external evidence is far too restricting. In order to judge the meaningfulness of any linking claim, the reader of any project report needs to be able to see the evidence of test validity presented in a systematic and theoretically sound manner. For this reason, we feel that the validation stage should present at least an overview of the evidence the developers feel show that the test is of sufficient quality and also that the linking claim is embedded in the test, and not simply related to two aspects of test validity. We also suggest that the validity evidence be based on a explicit model of validation such as that of Weir (2005).

The Exemplar Tasks

To date, the Council of Europe has recommended a small number of tasks and performances that are felt to reflect the different levels of the CEFR. These are said to be *standardised* to the levels. We strongly disagree that the tasks are in fact standardised in the technical sense, and suggest that the use of this term be dropped in any future manual edition. We believe that the tasks should be referred to only as *recommended* by the Council of Europe as being representative of the level. Until we get empirical evidence that the tasks have been standardised to the level they should remain as recommendations only.

City & Guilds will be presenting a small number of the reading and listening tasks and some writing tasks and exemplar performances of these tasks from Communicator to the Council of Europe as examples of the CEFR Level B2. We feel that all successful linking projects should be encouraged to do likewise is order to build up a more representative bank of exemplars at all CEFR Levels.

7.3.4. For the CEFR Itself

Rather than repeat the criticisms of the CEFR (e.g. Weir 2005b), we would like to focus instead on the fact that the relatively high level and quality of the descriptions of the different levels of the productive skills is not reflected in the descriptions of the receptive skills. While we had some difficulty with these skills at CEFR Level B2, there appears to be even greater problems at other levels (especially the C levels). The other issue we had was the apparent lack of any clearly stated model of reading or listening ability that drives the descriptions used in the CEFR. We feel that much work is needed in these areas if the CEFR is to be of real value in the future.

City & Guilds has done much work to try to clarify the grammatical forms that should appear at the different CEFR levels. We feel that this work needs to be taken on by other examination boards so that a clear understanding of broad syntactic development through the levels can be developed. Initiatives such as the English Profile Project in the UK mark one way of proceeding, though unless all of the major examination boards are involved there is some considerable likelihood that any resulting findings will represent a narrow perspective and may not add sufficiently to our understanding of the area.

7.3.5. Limitations

Like any project, this one was not without its limitations. Pressure of time and resources, for example limited the number of participants in the validation project, though the population for these tests was sufficiently large for us to make strong claims about the level and quality of the test.

Some readers might see the relatively small size of the expert panels will limit their value. We take a very different perspective. We feel that the large panels put in place by some other examination boards for their standard setting events add little to the quality of their decisions. A smaller panel of truly expert judges who are encouraged to debate and discuss their decisions are more likely to provide more valid evidence for the developer. We also feel that the panels should be comprised of members who represent both the developing institution and the broader testing community. This means that the claims that can be made after the standardisation stage will hold more meaning for test users.

7.4. Concluding Comments

In this project we set out to establish a link between the City & Guilds Communicator examination from their ESOL Suite and CEFR Level B2. We used the methodology suggested in the Pilot Manual (Council of Europe, 2003) as a basis for the project methodology, and found that some changes were needed to that methodology in order to ensure that the process described there reflected the reality of our experiences.

The process changed the whole way in which City & Guilds as an institution approaches the assessment of English proficiency. In fact, we should say that the process changed the institution. The decision to embed the philosophy of quality and the CEFR in the development and delivery of its examinations brought with it a substantial leap forward in the professionalization of the assessment practices of City & Guilds.

In many ways, we feel that this is the greatest strength of the movement towards benchmarking tests to the CEFR. By encouraging examination boards across Europe (and beyond) to use a systematic methodology to establish the quality and level of their test the Council of Europe has added to recent moves across the continent towards the localisation and professionalization of language test development. We feel that this should be seen by testing practitioners as a very positive movement.

Like the CEFR itself, the Pilot Manual is not without its flaws. The specification forms are at times vague and repetitive, the methodology itself can be interpreted as being linear in nature, and the expectations of the validation stage dated and limited. However, with more and more well designed and delivered linking projects taking place, the feedback received will only add to the value and usefulness of any later editions.

8. References

Berk, R. A. 1986. A consumer's guide to setting performance standards on criterion referenced tests. *Review of Educational Research*, 56, 137-172.

Chizek, G. J. and Bunch, M. B. 2007. *Standard Setting*. Thousand Oaks, CA: Sage.

City & Guilds. 2007. *Access to assessment and Qualifications: Guidance and regulations relating to candidates who are eligible for adjustments in assessment*. Downloaded, Sept. 10 2009 from http://www.cityandguilds.com/documents/ind_generic_docs_policydocs/Access_to_assessment_and_qualifications_v4.3.pdf

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

Council of Europe. 2003. *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*: *Preliminary Pilot Manual*. Strasburg: Council of Europe, Language Policy Division.

Educational Testing Services. 2007. *Mapping TOEFL iBT on the Common European Framework of Reference: Executive Summary*. Princeton, NJ: Educational Testing Service.

Impara, J.C & Plake, B. S. 1998. Teachers' ability to estimate item difficulty: a test of the assumptions I the Angoff standard setting method. *Journal of Educational Measurement*, 35, 69-81.

QALSPELL. 2004. *Quality Assurance in Language for Specific Purposes, Estonia, Latvia, Lithuania*. Leonardo da Vinci funded project. Website accessed 8 June 2008: http://www.qalspell.ttu.ee/.

Rethinasamy, S. 2006. *The effects on rating performance of different training interventions*. Unpublished PhD thesis, Roehampton University, London, UK.

Roehampton University and Universidad Veracruzana. 2008. *Exaver: affordable language test development project*. Website accessed 8 June 2008: http://www.uv.mx/exaver/nuv/index.html.

Schmitt, D. 2008. *Analysis of Reading Test Texts using Lexical Profiles and Coh-matrix*. Unpublished mimeo. Commissioned research by City & Guilds.

Tannenbaum, R. J. and Wylie, E. C. 2004. *Mapping test scores onto the Common European Framework: Setting standards of language proficiency on the Test of English as a Foreign Language (TOEFL), The Test of Spoken English (TSE), The Test of Written English (TWE), and The Test of English for International Communication (TOEIC)*. Princeton, NJ: Educational Testing Service.

Tannenbaum, R. J. and Wylie, E. C. 2005. *Mapping English Language Proficiency Test Scores Onto the Common European Framework*. Research Report 05-18. Princeton, NJ: Education Testing Services.

Tannenbaum, R. J. and Wylie, E. C. 2007. *Mapping TOEFL iBT, TOEIC, and TOEIC Bridge on to the Common European Framework: Interim Report*. Princeton, NJ: Education Testing Services.

Trinity College. Undated. *Relating the Trinity College London GESE and ISE examinations to the Common European Framework of Reference – project summary*. London: Trinity College London.

Weir, C. J. 2005a. *Language Testing and Validation: an evidence-based approach.* Oxford: Palgrave.

Weir, C. J. 2005b. Limitations of the Council of Europe's Framework of Reference (CEFR) in developing comparable examinations and tests. *Language Testing 22*(3), pages 282-300.

Lunz, Mary E., and Wright, Benjamin D. 1997. Latent Trait Models for Performance Examinations. In Jürgen Rost and Rolf Langeheine (Eds) *Applications of Latent Trait and Latent Class Models in the Social Sciences.* http://www.ipn.uni-kiel.de/aktuell/buecher/rostbuch/ltlc.htm

9. Appendices

Appendix 1 Communicator level – B2 Original Test Syllabus Overview with CEFR Linking Rationale

LISTENING SYLLABUS	CEFR criteria	Tested in part/s
1. understand standard spoken English delivered at normal speed	Can understand standard spoken language, live or broadcast, on both familiar and unfamiliar topics normally encountered in personal social, academic or vocational life. Only extreme background noise, inadequate discourse and/or idiomatic usage influences the ability to understand. (Page 66 Overall listening comprehension) Can understand in detail what is said to him/her in the standard spoken language even in a noisy environment. (Page 75 Understanding the native speaker interlocutor)	**1, 2, 3 & 4**
2. follow short conversations both formal and informal in a range of familiar situations understanding gist, context, purpose, function, attitude, feelings, opinions and relationships	Can follow the essentials of lectures, talks and reports and other forms of academic/professional presentations which are propositionally and linguistically complex. (Page 67 Listening as a member of a live audience) Can understand recordings in standard dialect likely to be encountered in social, professional or academic life and identify speaker view points and attitudes as well as the information content. (Page 68 Listening to audio media and recordings)	**2**
3. follow a conversation and predict the likely outcome	Can keep up with an animated conversation between native speakers. (Page 66 Understanding conversation between native speakers)	**2 & 4**
4. understand narratives, sequences, instructions, descriptions and explanations	Can understand detailed instructions reliably. (Page 79 Goal oriented co-operation)	**2 & 3**
5. identify the function of short utterances which may contain idiomatic/ expressions (See Functions and Grammar)		**1**
6. follow a discussion to identify gist, detail, purposes and key ideas and distinguish between fact and opinion	Can follow the discussion on matters related to his/her field, understand in detail the points given prominence by the speaker. (Page 78 Formal discussion and meetings)	**4**
7. extract and reproduce key information from announcements, media broadcasts, presentations and lectures including abstract and concrete topics encountered in personal, social, academic and vocational life	Can understand announcements and messages on concrete and abstract topics spoken in standard dialect at normal speed. (Page 67Listening to announcements and instructions) Can synthesise and report information and arguments from a number of sources. (Page 81 Information exchange) Can pass on detailed information reliably. (Page 81 Information exchange)	**3**

LISTENING SYLLABUS	CEFR criteria	Tested in part/s
8. follow clearly structured extended speech and more complex argument when familiar with the topic	Can follow extended speech and complex lines of argument provided the topic is reasonably familiar, and the direction of the talk is sign-posted by explicit markers. (Page 66 Overall listening comprehension) Can keep up with an animated discussion, identifying accurately arguments supporting and opposing points of view. (Page 78 Formal discussion and meetings)	**4**
Phonological features		
9. recognise how intonation, pitch and/or stress can affect meaning		**1, 2 & 4**
10. recognise feelings, moods, attitudes, important points and opinions expressed through stress, pitch and intonation	Can understand most radio documentaries and most other recorded or broadcast audio-material delivered in standard dialect and can identify the speaker's mood, tone etc. (Page 68 Listening to audio media and recordings)	**2 & 4**
Range		
2. understand ideas, arguments and descriptions expressed through complex sentence forms	Can understand the main ideas of propositionally and linguistically complex speech on both concrete and abstract topics delivered in standard dialect, including technical discussions in his/her filed of specialisation. (Page 66 Overall listening comprehension)	**3 & 4**
12. understand some lower frequency vocabulary and expressions relating to everyday life and current events		**1, 3 & 4**
Register		
6. recognise degrees of formality used by speakers in different types of utterances in everyday and less familiar situations		**1, 2 & 4**
Understanding gist		
7. understand the main ideas in longer but clearly structured announcements , conversations and discussions on familiar and unfamiliar concrete and abstract topics	Can understand the main ideas of propositionally and linguistically complex speech on both concrete and abstract topics delivered in standard dialect, including technical discussions in his/her filed of specialisation. (Page 66 Overall listening comprehension)	**3 & 4**
Understanding detail		
8. extract the more salient points of detail from longer but clearly structured texts on familiar and unfamiliar topics and on both concrete and abstract topics.		**3 & 4**

Reading syllabus	CEFR criteria	Tested in part/s
1. understand texts in different styles and purposes with a large degree of independence	Can read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively. (Page 69 Overall reading comprehension)	**1, 2, 3 & 4**
2. understand the main ideas in complex texts on both familiar and abstract topics	Can understand specialised articles outside his/her field, provided he/she can use a dictionary occasionally to confirm his/her interpretation of terminology. (Page 70 Reading for information and argument)	**4**
3. understand the way meaning is built up in a range of texts		**1, 2 & 3**
4. locate specific information from different parts of a text or different texts	Can scan quickly through long and complex texts, locating relevant details. (Page 70 Reading for orientation) Can obtain information, ideas and opinions from highly specialised sources within his/her field. (Page 70 Reading for information and argument)	**3 & 4**
5. understand feelings, opinions, warnings and conditions in both formal and informal text		**3 & 4**
6. understand lengthy texts containing complex instructions or explanations	Can understand lengthy complex instructions in his/her field, including details on conditions, warnings, provided he/she can reread difficult sections. (Page 71 Reading instructions)	**4**
7. understand articles and reports concerned with contemporary issues in which the writers adopt particular viewpoints	Can understand articles and reports concerned with contemporary issues in which the writers adopt particular viewpoints. (Page 70 Reading for information and argument) Can quickly identify the content of news items, articles and reports on a wide range of professional topics, deciding whether closer study is worthwhile. (Page 70 Reading for orientation)	**4**
8. locate and understand information, ideas and opinions from longer more specialised sources in familiar contexts	Can obtain information, ideas and opinions from highly specialised sources within his/her field. (Page 70 Reading for information and argument)	**4**
Range		
9. understand a broad range of vocabulary but may experience some difficulty with low-frequency idioms	Has a broad active reading vocabulary, but may experience some difficulty with low frequency idioms. (Page 69 Overall reading comprehension)	**1, 2, 3 & 4**
10. understand texts which contain a broad range of grammatical structures		**1, 2, 3 & 4**
Register		
11. understand the features of register in texts including those conveying emotion or dispute		**2 & 3**
Text structure		

Reading syllabus	CEFR criteria	Tested in part/s
12. recognise how purpose is achieved in a range of texts including those containing images, graphical and tabular data		**1, 2, 3 & 4**
13. understand a broad range of discourse markers including those expressing addition, cause and effect, contrast, sequence and time		**1, 2, 3 & 4**

Writing syllabus	CEFR criteria	Tested in part/s
1. write coherently on topics of general interest linking ideas appropriately and effectively	Can write clear, detailed texts on a variety of subjects related to his/her field of interest, synthesising and evaluating information and arguments form a number of sources. (Page 61 Overall written production) Can write news and views effectively in writing and relate to those of others. (Page 83 Overall written interaction)	**1 & 2**
2. write clear connected text describing real or imaginary people or events	Can write clear, detailed descriptions of real or imaginary events and experiences, marking the relationship between ideas in clear connected text, and following established conventions of the genre concerned. (Page 62 Creative writing)	**1 & 2**
3. present an argument giving points for and against, supporting and evaluating different views	Can write an essay or report which develops an argument giving reasons in support of or against a particular point of view and explaining the advantages and disadvantages of various options. (Page 62 Reports and essays) Can evaluate different ideas or solutions to problems. (Page 62 Reports and essays)	**1 & 2**
4. write formal letters, reports or articles to fulfil a range of functions for practical purposes	Can synthesise information and arguments form a number of sources. (Page 62 Reports and essays) Can pass on detailed information reliably. (Page 129 Propositional precision)	**1 & 2**
5. write letters, descriptions of personally significant events, people or experiences	Can write clear detailed descriptions on a variety of subjects related to his/her field of interest. (Page 62 Creative writing) Can write letters conveying degrees of emotion and highlighting the personal significance of events and experiences commenting on the correspondent's news and views. (Page 83 Correspondence) Can develop a clear description or narrative, expanding and supporting his/her main points with relevant detail and examples. (Page 125 Thematic development)	**1 & 2**
Accuracy		
6. use correct punctuation in formal and informal writing to enhance meaning	Spelling and punctuation are reasonably accurate but may show signs of mother tongue influence. (Page 118 Orthographic control)	**1 & 2**
7. correctly spell words used in work, study and daily life	Lexical accuracy is generally high, though some confusion and incorrect word choice does occur without hindering communication. (Page 112 Vocabulary control)	**1 & 2**
8. control grammar to communicate effectively although errors may occur when complex structures are attempted	Shows a relatively high degree of grammatical control. Does not make mistakes which lead to misunderstanding. (Page 114 Grammatical accuracy)	**1 & 2**
Range		

Writing syllabus	CEFR criteria	Tested in part/s
9. use words and expressions appropriate to topic and purpose of the writing	Has a good range of vocabulary for matters connected to his/her field and most general topics. (Page 112 Vocabulary range) Can express him/herself clearly without much sign of having to restrict what he/she wants to say. (Page 110 General linguistic range) Has a sufficient range of language to be able to give clear descriptions, express viewpoints and develop arguments without much conspicuous searching for words, using some complex sentence forms to do so. (Page 110 General linguistic range)	**1 & 2**
10. adjust register in familiar contexts to suit purpose and readership	Can adjust what he/she says and the means of expressing it to the situation … (Page 124 Flexibility)	**1 & 2**
Organisation		
11. use a range of linking words effectively to show clearly the relationship between ideas	Can use a variety of linking words effectively to mark the relationship between ideas. (Page 125 Coherence and cohesion)	**1 & 2**
12. paragraph appropriately	Can produce clearly intelligible continuous writing which follows standard layout and paragraphing conventions. (Page 118 Orthographic control)	**1 & 2**
13. reproduce conventional features of common types of text		**1 & 2**

Speaking syllabus	CEFR criteria	Tested in part/s
1. speak with a degree of fluency and spontaneity making sustained interaction possible without undue strain	Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without imposing strain on either party. (Pages 74 Overall spoken interaction & 129 Spoken fluency)	**1, 2, 3 & 4**
2. communicate personal information, opinions, feelings and ideas	Can express his/her ideas and opinions with precision, present and respond to complex lines of argument convincingly. (Page 78 Formal discussion and meetings)	**1, 3 & 4**
3. communicate in a variety of social situations using a range of functional language		**2**
4. exchange information to perform a task	Can understand and exchange complex information and advice on a full range of matters related to his/her occupational role. (Page 81 Information exchange) Can give a clear, detailed description of how to carry out a procedure. (Page 81 Information exchange)	**3**
5. narrate, describe, explain and express opinions in extended speech	Can give clear, detailed descriptions and presentations on a wide range of subjects related to his/her field of interest, expanding and supporting ideas with subsidiary points and relevant examples. (Pages 58 Overall oral production)	**3 & 4**
6. give straightforward descriptions, narratives, directions, instructions on topics encountered in personal, social, academic or vocational life	Can give clear, detailed descriptions on a wide range of subjects related to his/her field. (Page 59 Sustained monologue). Can use the language fluently, accurately and effectively on a wide range of general, academic, vocational or leisure topics, marking clearly the relationships between ideas. (Page 74 Overall spoken interaction) Can develop a clear description or narrative, expanding and supporting his/her main points with relevant detail and examples. (Page 125 Thematic development)	**4**
7. contribute points to an argument on a familiar topic integrating sub-themes and coming to a conclusion.	Can contribute, account for and sustain his/her opinion, evaluate alternative proposals and make and respond to hypotheses. (Page 78 Formal discussion and meetings)	**3 & 4**
Pronunciation		
8. pronounce clearly the sounds of English in connected speech	Has acquired a clear, natural pronunciation and intonation. (Page 117 Phonological control)	**1, 2, 3 & 4**

Speaking syllabus	CEFR criteria	Tested in part/s
9. produce stretches of language with few noticeable long pauses, but with some hesitation when searching for patterns and expressions	Can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he/she searches for patterns and expressions, there are few noticeable pauses. (Page 129 Spoken fluency)	**1, 2, 3 & 4**
Accuracy		
10. display a relatively high degree of grammatical control without impeding errors	Can communicate spontaneously with good grammatical control without much sign of having to restrict what he/she wants to say, adopting a level of formality appropriate to the situation. (Page 74 Overall spoken interaction & 129 Spoken fluency)	**1, 2, 3 & 4**
Range		
11. use sufficient range of language to give detailed descriptions and arguments and be able to highlight personal events and emotions	Can highlight the personal significance of events and experiences, account for and sustain views by clearly by providing relevant explanations and arguments. (Page 74 Overall spoken interaction) Can convey degrees of emotion and highlight… (Page 76 Conversation)	**1, 2, 3 & 4**
12. produce complex sentences although there is still some searching for vocabulary and expressions	Has sufficient range of language to be able to give clear descriptions, express viewpoints and develop arguments without much conspicuous searching for words, using complex sentence forms to do so. (Page 110 General linguistic range)	**1, 2, 3 & 4**
Register		
13. adopt a degree of formality appropriate to the circumstances	…and adopt a level of formality appropriate to the situation. (Page 74 Overall spoken interaction& 124 Flexibility) Can express him/herself confidently, clearly and politely in a formal or informal register, appropriate to the situations and person(s) concerned. (Page 122 Sociolinguistic appropriateness)	**1, 2, 3 & 4**
14. cope linguistically with more stressful kinds of interaction such as complaints or disputes	Can cope linguistically to negotiate a solution to a dispute like an undeserved traffic ticket, financial responsibility for damage to a flat, for blame regarding an accident. (Page 80 Transaction to obtain goods and services)	**2**
Fluency		

Speaking syllabus	CEFR criteria	Tested in part/s
15. manage the conventions of turn taking using appropriate phrases for making and dealing with interruptions and requesting information	Can initiate, maintain and end discourse appropriately with effective turn taking. (Pages 86 Taking the floor & 124 Turntaking). Can initiate discourse, take his/her turn when appropriate and end conversations when he/she needs to, though may not always do this elegantly. (Pages 86 Taking the floor & 124 Turntaking)	**2 & 3**
16. link utterances using some cohesive devices although there may be some 'jerkiness' in extended speech	Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse, though there many be some jumpiness in a long contribution. (Page 125 Coherence and cohesion)	**1 & 4**

CEFR DRAFT LINKING MANUAL

SPECIFICATION FORMS FOR COMMUNICATOR LEVEL

B2

June 2008

GENERAL EXAMINATION DESCRIPTION	
General Information Name of examination	International ESOL - (Preliminary, Access, Achiever, Communicator, Expert, Mastery) All 6 levels are included on these forms. Where necessary differences in the levels are indicated
Language tested	English
Examining institution	City & Guilds
Date of this version	November 2006
Type of examination	☑ International ☐ National ☐ Regional ☐ Institutional
Purpose	To test general proficiency in English – focusing on reading, writing and listening skills (there is a separate paper for Speaking which is not included in this specification)
Target population	☑Lower Sec ☑Upper Sec ☑Uni/College Students ☑Adult
No. of test-takers per year	50,000 + expected

What is the overall aim?

To allow the test user to draw inferences, based on test performance, that the level of proficiency of the successful candidate is at CEFR level B2 in the areas of reading, writing and listening

What are the more specific objectives? If available describe the needs of the intended users on which this examination is based.

This suite of examinations is targeted specifically at non-native speakers of English (young people & adults) worldwide who require:

a) recognized certification of the level of their English language competence in reading, writing and listening
b) a series of graded examination to provide steps up the ladder of proficiency of English

These English examinations can be stand alone or taken as a complement to the City & Guilds International Spoken ESOL Suite.

What is/are principal domain(s)?	☑Public ☑Personal ☐Occupational ☑Educational	
Which communicat-ive activities are tested?	☑ **1** Listening comprehension ☑ **2** Reading comprehension ☐ **3** Spoken interaction ☑**4** Written interaction ☐ **5** Spoken production ☑ **6** Written production ☐ **7** Integrated skills ☐ **9** Spoken mediation of text ☐ **10** Written mediation of text ☑ **11** Language (e.g. Grammar, Vocabulary, Cohesion) ☐ **12** Other: (specify):	***Name of Subtest(s)*** 1. Listening 2. Reading 3. Writing

Give name and duration of test subtests (referred to above right):	Name of Examination: International ESOL Name of Subtest 1 Listening 2 Reading 3 Writing	Duration of Examination(s) Communicator: 2.5 hours No specified duration for each subtest
What type(s) of test tasks are used?		Subtests used in (Write numbers above)
	☑ Multiple choice ☐ True/False ☑ Matching ☐ Ordering ☐ Gap fill sentence ☐ Sentence completion ☑ Gapped text / cloze, selected response ☑ Open gapped text / cloze / ☑ Short answer to open question(s) ☑ Extended answer (text / monologue) ☐ Interaction with examiner ☐ Interaction with peers OTHER ☑ Information transfer	1 2 2 3 3 3 1, 2
What Information is published for candidates and teachers?	☑ Overall aim ☑ Principal domain(s) ☑ Test subtests ☑ Test tasks ☑ Sample test papers ☐ Video of format of oral	☑ Sample answer papers ☑ Marking schemes ☐ Grading schemes ☑ Standardised performance samples showing pass level ☐ Sample certificate
What is Reported?	☑ Global Grade ☑ Grade per subtest	☐ Global Grade plus graphic profile ☑ Profile per subtest ☑

Form A1: General Examination Description

Test development	**IESOL**
What organisation decided that the examination was required?	☑ Own organisation/school ☐ A cultural institute ☐ Ministry of Education ☐ Ministry of Justice
If an external organisation is involved, what influence do they have on design and development?	☐ Determine the overall aims ☐ Determine level of language proficiency ☐ Determine examination domain or content ☐ Determine exam format and type of test tasks
If no external organisation was involved, what other factors determined design and development of examination?	☑ A needs analysis – *informal, qualitative feedback was collected from experts and centres on the previous suite of Pitman ESOL examinations* ☑ Internal description of examination aims ☑ Internal description of language level *Based on CEFR – see Appendix 1* ☑ A syllabus or curriculum ☑ Profile of candidates
In producing test tasks are specific features of candidates taken into account? *Every effort is taken to ensure that the questions included are free from bias and are in line with the guidelines of: APA (American Psychological Association) standards of educational and psychological testing, AERA, NAPA and NCME.*	☑ Linguistic background (L1) ☑ Language learning background ☑ Age ☑ Educational level ☑ Socio-economic background ☑ Social-cultural factors ☑ Ethnic background ☑ Gender
Who writes the items or develops the test tasks?	A contracted team of expert item writers – trained to City & Guilds standards and experienced working with CEFR
Have test writers guidance to ensure quality?	☑Training ☑Guidelines ☑ Checklists ☑Examples of valid, reliable, appropriate tasks: ☑ Calibrated to CEFR level description ☐ Calibrated to other level description:
Is training for test writers provided?	☑ Yes initially with City & Guilds/CEFR specifications, then ongoing feedback on item performance & routine in-service training ☐ No

Are test tasks discussed before use?	☑ Yes ☐ No
If yes, by whom?	☑ Individual colleagues ☑ Internal group discussion ☑ External examination committee ☑ Internal stakeholders ☐ External stakeholders
Are tests tasks pre-tested?	☑ Yes ☐ No
If yes, how?	With a cohort of candidates who have been identified by centres as being at the appropriate level of the examination being pretested. Where candidate numbers permit, reading and listening tasks are pretested with > 100 candidates For the writing tasks qualitative data from the interlocutors and exam centres are collected on the suitability and functionality of the tests, which helps the Vetting Committee Examination Review Committee establish the validity of the tasks
If no, why not?	
Is the reliability of the test estimated?	☑ Yes ☐ No
If yes, how?	☑ Data collection and psychometric procedures – Cronbach's alpha ☑ Other: Scorer Reliability (inter- & intra-rater) is calculated for the rating of the writing tasks
Are different aspects of validity estimated?	☐ Face validity – during piloting - questionnaires to teachers in examination centres ☑ Content validity (ensured by providing detailed Item and Editor Specifications & obtaining advice from external consultants). Undertaken as part of the re-specification of the test for this project. ☑ Concurrent validity – (during development this was done for two levels and correlated with Item banker). This CEFR linking project is providing evidence of concurrent validity in relation to test level. ☐ Predictive validity ☑ Construct validity inter subtest correlations of the sub tests
If yes, describe how?	(See above) By a team of trained experts during the

	development process, analysis by external consultants and qualitative feedback during the pilot phase.

Form A2:Test Development

Marking: Subtest	***IESOL***
How are the test tasks marked?	For receptive test tasks: ☐ Optical mark reader ☑ Clerical marking For productive or integrated test tasks: ☑ Trained examiners ☐ Teachers
Where are the test tasks marked?	☑ Centrally ☐ Locally: ☐ By local teams ☐ By individual examiners
What criteria are used to select markers?	Potential markers must be experienced ESOL teachers. They need to have had training on both City & Guilds IESOL examinations, as well as familiarisation training on the CEFR. Performance on examinations monitored.
How is accuracy of marking promoted?	☑ Regular checks by co-ordinator ☑ Training of markers/raters ☑ Moderating sessions to standardise judgements ☑ Using standardised examples of test tasks: ☑ Calibrated to CEF –levels ☑ Calibrated to City & Guilds Levels ☐ Not calibrated to CEF or other description
Describe the specifications of the rating criteria of productive and/or integrative test tasks.	☑ one holistic score for each task ☐ marks for different aspects for each task ☐ rating scale for overall performance in test ☑ rating grid for aspects of test performance ☐ rating scale for each task ☐ rating grid for aspects for each task ☐ rating scale bands are defined, but not to CEFR ☑ rating scale bands are defined in relation to CEFR
Are productive or integrated test tasks single or double rated?	☐ Single rater ☐ Two simultaneous raters ☑ Double marking of scripts (fixed date exams) ☑ Other: specify: non fixed date exams are single rated (examiners are monitored through regular standardisation and moderation – this system is under review with systematic double marking to be introduced)

If double rated, what procedures are used when differences between raters occur?	☑ Use of third rater and that score holds ☐ Use of third marker and two closest marks used ☐ Average of two marks ☑ Two markers discuss and reach agreement Initially the markers use social moderation to reach agreement. If they are unable to reach a decision then the team leader (as third marker) decides and that score holds
Is inter-rater agreement calculated?	☑ Yes measuring the inter-rater reliability currently with Spearman Rho – from 2008 Multi-faceted Rasch analysis used to estimate inter and intra-rater reliability ☐ No

Form A3: Marking

Grading:	**IESOL**
Are pass marks and/or grades given?	☑ Pass marks ☑ Grades
Describe the procedures used to establish pass marks and/or grades and cut scores	The pass mark is set using a modified Angoff standard setting procedure. Judges are asked to estimate the cut score based on a definition of the minimally competent candidate at level B2 (defined for each skill area prior to the event). Data are used to support the judgements (made over two rounds). The cut score for a First Class Pass is set at a point where the candidate can be said to have met some of the criteria for the next highest level.
If grades are given, how are the grade boundaries decided?	The cut off scores have been provisionally calculated by expert validation and confirmed with a concurrent validity test using a test devised from Eurocentres Itembanker. The level of the cut off scores will continue to be monitored through ongoing analysis of the live tests. These cut scores have been supported by the standard setting element of this project.
If only pass / fail is reported, How are the cut-off scores for pass / fail set?	See above
How is consistency in these standards maintained?	Consistency is maintained by ensuring that the parallel versions of the test are equivalent. Item writers are carefully trained and follow the guidelines laid down in the Item Writers' Guide. In addition there is a Vetting Panel and an Examination Review Committee who also check the content validity and the results of the statistical analyses carried out during the pretesting phase. Consistency in the marking is maintained through training, double marking, standardisation and moderation.

Form A4: Grading

Results	IESOL
What results are reported to candidates?	☑ Global grade or pass / fail ☑ Grade or pass / fail per subtest ☐ Global grade plus profile across subtests ☑ Profile of aspects of performance per subtest– (a description on aspects of candidate performance using a Performance Code Report - initially only for failing candidates, but from 2007 available for all candidates)
In what form are results reported?	☑ Raw scores for the listening and reading subtest initially available for failing candidates only, but since 2007 also available for pass candidates ☑ Undefined grades (e.g. “C”) ☐ Level on a defined scale ☑ Diagnostic profiles (see above comments)
On what document are results reported?	☐ Letter or email ☑ Report card ☑ Certificate / Diploma
Is information provided to help candidates to interpret results? Give details.	A key is provided for the performance codes used in the diagnostic reports for failing candidates. Additional details available from 2008 on the dedicated website on the meaning of the CEFR levels that are used on the certificate – based on ‘Can Do’ statements.
Do candidates have the right to see the corrected and scored examination papers?	Yes
Do candidates have the right to ask for remarking?	Yes, for a fee

Form A5: Reporting Results

Test Analysis and Post-examination Review	IESOL
Is feedback gathered on the examinations?	☑ Yes, in the course of pre-testing & live testing ☐ No
If yes, by whom?	☑ Internal experts (colleagues) ☑ External experts ☑ Local examination institutes ☑ Test administrators ☑ Teachers ☑ Candidates
Is the feedback incorporated in revised versions of the examinations?	☑ Yes ☐ No
Is data collected to do analysis on the tests?	☑ On all tests (all tasks pretested with samples of up to 100. Data also collected on live tests) ☐ No
If yes, indicate how data are collected?	☑ During pre-testing ☑ During live examinations ☐ After live examinations
For which features is analysis on the data gathered carried out?	☑ Difficulty ☑ Discrimination ☑ Reliability ☑ Validity
State which analytic methods have been used (e.g. in terms of psychometric procedures).	• Descriptive stats - measures of central tendency and dispersion • Classical item statistics • IRT – item level difficulty and item misfit • Qualitative feedback (how it works/rater remarks) • Inter-subtest correlations
Are performances of candidates from different groups analysed?. If so, describe how.	Yes –bias analysis based on the candidate data performed during annual test review
Describe the procedures to protect the confidentiality of data.	All scripts are handled and stored within secure areas. Data are analysed using spreadsheets held on a secure network drive. There is limited access to this data.
Are relevant measurement concepts explained for test users? If so, describe how.	Summary of how final scores are calculated is available

Form A6: Data Analysis

Rationale for making decisions	IESOL
Give the rationale for the decisions that have been	The main objective in the development

made in relation to the examination or the test tasks in question.	of the International ESOL qualifications was to produce a test that is: **Valid in terms of content** That it follows the test specification & task guidelines laid out in the Item Writer and Editor Specifications. **At the appropriate level of the CEF** i.e. levels can be interpreted in terms of CEF scales **Reliable** Test specifications are drawn up so as to ensure as far as possible that the level is maintained from one test form to another. **Expected to produce positive washback** **User friendly** The tasks are designed to be accessible to candidates with a wide range of educational backgrounds; to test language ability in direct ways without placing unnecessary additional cognitive burdens on the candidate. **Practical** in terms of administration

Form A7: Rationale for Decisions

Initial Impression of Overall CEF Level		
☐ **A1** ☐ **A2**	☐ **B1** ☑ **B2 = Communicator**	☐ **C1** ☐ **C2**

Short rationale, reference to documentation

The assessment standards for the International ESOL (IESOL) examination were drawn up from the CEFR illustrative frameworks.

The assessment syllabus for the subtests: reading, writing and listening are published on the following pages of the IESOL Handbook: preliminary 29-31, Access 41-43, Achiever 57-59, ***Communicator 76-78***, Expert 94-96, Mastery 109-111 (See Appendix 1 for an example at Communicator level). These show what a candidate is expected to do at each level and were developed from the CEF scales for communicative language activities in chapter 4 of the CEF.

Also stated on the above pages of the handbook are *the performance standards* that a candidate at each level is expected to reach. These standards were used:

- to inform the assessment criteria for International ESOL writing tasks;
- to devise the level of the items for the reading and listening sections;
- for creating the lists of grammar and functions needed for each level.

These **performance standards** have been developed from the scales for aspects of language proficiency in Chapter 5 of the CEF.

All the stakeholders involved in the production of these examinations have undergone a process of training, familiarisation and benchmarking with the CEF according to the recommendations laid down in the CEF Manual.

Form A8: Impression of Overall Examination Level

	IESOL LISTENING
Which situations, content categories, domains are the test takers expected to show ability in?	Domains: personal, public and educational.
Which communication themes are the test takers expected to be able to handle?	Self and Family, Home, Local area, Everyday life, Education, Free time Leisure interests, Entertainment, Travel, Relationships, Health and hygiene, Shopping, Food & drink, Public services, Places, Language, Weather, Measures and shapes.
Which communicative tasks are the test takers expected to be able to handle?	The communicative tasks are listed in the "Function lists" given for each level in the International ESOL handbook. The main headings for Communicator can be found on pages 91-93.
What kind of communicative activities and strategies are the test takers expected to be able to handle?	• listening to public announcements (information, instructions, warnings, etc.); • listening to media (radio, TV, recordings, cinema); • listening to overheard conversations, etc. In each case the user may be listening: • for gist; • for specific information; • for detailed understanding; • for implications, etc.
What text-types and what length of text are the test takers expected to be able to handle?	**Media**: audiotape / cassette or disc **Test Type**: • Interpersonal dialogues and conversations • Discussions • Public announcements and instruction • Broadcasts • Public speeches, lectures and presentation • Telephone conversation / and leaving messages Maximum text length is approx. 550 words.
What kind of tasks are the test takers expected to be able to handle?	• Identifying the communicative function of short utterances • Listening to identify specific aspect of a spoken dialogue / conversations • Extracting specific detail information from spoken explanations, messages and announcements • Listening for gist, opinion and speaker's attitude in discussions
After reading the scale for Overall Listening	**Level: B2**

Comprehension, given below, indicate and justify at which level(s) of the scale the subtest should be situated.	***Justification*** ***(incl. reference to documentation)*** The assessment standards, exam tasks, input texts and assessment criteria for the International ESOL examination were drawn up from the CEF. Can understand the main ideas of propositionally and linguistically complex speech on both concrete and abstract topics delivered in a standard dialect, including technical discussions in his/her field of specialisation. Can follow extended speech and complex lines of argument provided the topic is reasonably familiar, and the direction of the talk is sign-posted by explicit markers. For more evidence see: International ESOL Handbook Item Writer & Editor Guide Marking Guide

Form A9: Listening Comprehension

	IESOL READING
Which situations, content categories, domains are the test takers expected to show ability in?	Domains: personal, public and educational.
Which communication themes are the test takers expected to be able to handle?	Self and Family, Home, Local area, Everyday life, Education, Free time Leisure interests, Entertainment, Travel, Relationships, Health and hygiene, Shopping, Food & drink, Public services, Places, Language, Weather, Measures and shapes.
Which communicative tasks are the test takers expected to be able to handle?	The communicative tasks are listed in the "Function lists" given for each level in the International ESOL handbook. The main headings for Communicator can be found at pages 91-93.
What kind of communicative activities and strategies are the test takers expected to be able to handle?	**Communicative activities** • reading for general orientation; • reading for writer's perspective; • reading for information (correspondence, magazine articles, instructions etc.) • reading and following instructions; The language user may read: • for gist; • for specific information; • for detailed understanding; • for implications, etc.
What text-types and what length of text are the test takers expected to be able to handle?	**Text Types:** address, advertisement, appointment card, bill, brochure, calendar, cheque, diary, form, greetings card, guide, informative article, instructions, label, leaflet, legend, list, menu, message, note, notice, correspondence, poster, price-list, product packaging, radio/theatre/TV programme, recipe, record, signs, short biography, table, telephone directory, ticket, timetable, weather forecast. Maximum text length for reading comprehension is 400 words
What kind of tasks are the test takers expected to be able to handle?	**Tasks:** • Complete texts (variety of text types) by inserting missing sentences / words into phrases • Retrieve specific information (range of texts or from an extended text) • Identify the purpose / function of texts • Scanning four short authentic texts • Locate and transfer specific information

After reading the scale for Overall Reading Comprehension, given below, indicate and justify at which level(s) of the scale the subtest should be situated.	**Level B2** ***Justification*** ***(incl. reference to documentation)*** The assessment standards, exam tasks, input texts and assessment criteria for the International ESOL examination were drawn up from the CEF. Can read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively. Has a broad active reading vocabulary, but may experience some difficulty with low frequency idioms. For more evidence see: International ESOL Handbook Item Writer & Editor Guide Marking Guide

Form A10: Reading Comprehension

	IESOL WRITTEN INTERACTION
Which situations, content categories, domains are the test takers expected to show ability in?	Domains: personal, public and educational.
Which communication themes are the test takers expected to be able to handle?	Communication Themes Covered (Found in the International ESOL Handbook) **Communicator Levels**: Self and Family, Home, Local area, Everyday life, Education, Free time Leisure interests, Entertainment, Travel, Relationships, Health and hygiene, Shopping, Food & drink, Public services, Places, Language, Weather, Measures and shapes.
Which communicative tasks are the test takers expected to be able to handle?	The list of communicative tasks are contained in the "Function lists" given for each level in the International ESOL handbook. The main headings for Communicator can be found on pages 91-93.
What kind of communicative activities and strategies are the test takers expected to be able to handle?	**Communicative Activities at all levels**: • Awareness of audience in all writing tasks
What kind of texts and text-types are the test takers expected to be able to handle?	**Text Types**: • Postcards • Magazine / newspaper articles
What kind of tasks are the test takers expected to be able to handle?	**Task Types:** • Write an informal letter, message, postcard or instructions • Write formal letters, instructions & reports • May also included discursive topics
After reading the scale for Overall Written Interaction, given below, indicate and justify at which level(s) of the scale the subtest should be situated.	**Level A2** ***Justification*** ***(incl. reference to documentation)*** The assessment standards, exam tasks, input texts and assessment criteria for the International ESOL examination were drawn up from the CEF. Can express news and views effectively in writing, and relate to those of others. For more evidence see: International ESOL Handbook Item Writer & Editor Guide Marking Guide

Form A12: Written Interaction

	IESOL WRITTEN PRODUCTION
Which situations, content categories, domains are the test takers expected to show ability in?	Domains: personal, public and educational.
Which communication themes are the test takers expected to be able to handle?	Communication Themes Covered (Found in the International ESOL Handbook) **Communicator Levels**: Self and Family, Home, Local area, Everyday life, Education, Free time Leisure interests, Entertainment, Travel, Relationships, Health and hygiene, Shopping, Food & drink, Public services, Places, Language, Weather, Measures and shapes.
Which communicative tasks are the test takers expected to be able to handle?	The communicative tasks are in the "Function lists" given for each level in the International ESOL handbook. The main headings for each level are as follows: • Personal environment, • Expressing thoughts and feelings, • Making things happen, • Social contact • Expressing attitudes
What kind of communicative activities and strategies are the test takers expected to be able to handle? **Note:** in original tables, CEFR 4.4.1.2 is suggested here (that information is for spoken interaction)	**Communicative Activities** • correspondence by letter, fax, e-mail, etc.; • writing informative texts • writing essays
What kind of texts and text-types are the test takers expected to be able to handle?	**Text Types:** • Descriptive • Narrative • Argumentative • Discursive
What kind of tasks are the test takers expected to be able to handle?	**Tasks:** • Write a report, article or instructions
After reading the scale for Overall Written Production, given below, indicate and justify at which level(s) of the scale the subtest should be situated.	**Level B2** ***Justification*** ***(incl. reference to documentation)*** The assessment standards, exam tasks, input texts and assessment criteria for the International ESOL examination were drawn up from the CEF. Can write clear, detailed texts on a variety of subjects related to his/her field of interest, synthesising and evaluating information and arguments from a number of sources. For more evidence see:

	International ESOL Handbook Item Writer & Editor Guide Marking Guide

Form A14: Written Production

LINGUISTIC COMPETENCE	
What is the range of lexical and grammatical competence that the test takers are expected to be able to handle?	For a detailed listing of the grammatical competence expected at each level please see the International ESOL Qualification Handbook. For Communicator see pages 79-86. The range of vocabulary required at each level is set by ensuring the input texts are at the appropriate level of the CEF. The candidates' lexical competence is then tested indirectly by their ability to complete the reading and listening tasks.
After reading the scale for Linguistic Competence in Table 4.3, indicate and justify at which level(s) of the scale the examination should be situated.	**Level B2** ***Justification*** ***(incl. reference to documentation)*** The assessment standards, exam tasks, input texts and assessment criteria for the International ESOL examination were drawn up using the CEF. Shows a relatively high degree of grammatical control Does not make mistakes which lead to misunderstanding. For more evidence see: International ESOL Handbook Item Writer & Editor Guide Marking Guide

Form A19a: Aspects of Linguistic Competence in Reception

SOCIO-LINGUISTIC COMPETENCE	
What are the socio-linguistic competences that the test takers are expected to be able to handle: linguistic markers politeness conventions, register, adequacy, dialect/accent, etc?	**Sociolinguistic Competence** The test takers are required to display an awareness of the sociolinguistic dimensions in the completion of both the listening and reading tasks that they are engaged in. The chief aspects of sociolinguistic competence include: candidate awareness of social conventions, relationships and politeness. These sociolinguistic competences are captured in the following tasks : Listening task 1 & 2 & Reading task 3
After reading the scale for Socio-linguistic Competence in Table 4.3, indicate and justify at which level(s) of the scale the examination should be situated.	**Level B2** ***Justification*** ***(incl. reference to documentation)*** The assessment standards, exam tasks, assessment criteria for the International ESOL examination were drawn up from the CEF. Can with some effort keep up with and contribute to group discussions even when speech is fast and colloquial. Can sustain relationships with native speakers without intentionally amusing or irritating them or requiring them to behave other than they would with a native speaker. Can express him or herself appropriately in situations and avoid crass errors of formulation. For more evidence see: * International ESOL Handbook * Item writer's and editor's checklist * Assessment Criteria in the Marking Guide

Form A19b: Aspects of Socio-Linguistic Competence in Reception

PRAGMATIC COMPETENCE	
What are the pragmatic competences that the test takers are expected to be able to handle: discourse competences, functional competences?	**Pragmatic Competences:** **Discourse Competence** Candidates are expected to display the discourse competence of **coherence and cohesion** in their ability to complete aspects of the following tasks: Reading Task 1 & Reading Task 2 **Functional Competences**- the candidates are expected to handle the "Function lists" given for each level in the International ESOL handbook. The main headings for Communicator can be found on pages 91-93.
After reading the scale for Pragmatic Competence in Table 4.3, indicate and justify at which level(s) of the scale the examination should be situated.	**Level B2** **Justification (incl. reference to documentation)** The assessment standards, exam tasks, input texts and assessment criteria for the International ESOL examination were drawn up from the CEF. Can use a limited number of cohesive devices to link his/her utterances into clear coherent discourse, though there maybe some jumpiness in along contribution. For more evidence see: International ESOL Handbook Item writer's and editor's checklist Assessment Criteria in the Marking Guide

Form A19c: Aspects of Pragmatic Competence in Reception

Strategic Competence	
What are the strategic competences that the test takers are expected to be able to handle?	Strategic competences for receptive activities are tested indirectly by the candidate's ability to "identify cues & infer" meaning in the reading and listening texts. By employing these strategies the candidate's likelihood of success in those tasks are directly improved.
After reading the scale for Strategic Competence in Table 4.3, indicate and justify at which level(s) of the scale the examination should be situated.	**Level B2** ***Justification*** ***(incl. reference to documentation)*** The assessment standards, exam tasks, input texts and assessment criteria for the International ESOL examination were drawn up from the CEF. Is skilled at using contextual, grammatical and lexical cues to infer attitude, mood and intentions and anticipate what will come next. The ability to infer is a key objective tested in the reading test. For more evidence see: International ESOL Handbook Item writer's and editor's checklist Assessment Criteria in the Marking Guide

Form A19c: Aspects of Strategic Competence in Reception

LINGUISTIC COMPETENCE	
What is the range of lexical and grammatical competence that the test takers are expected to be able to handle?	For a detailed listing of the grammatical competence expected at each level please see the International ESOL Qualification Handbook pp 79-86 Lexical competence in Written Interaction is captured in the assessment criteria of "**Range**"
What is the range of phonological and orthographic competence that the test takers are expected to be able to handle?	Familiarity with the Roman alphabet and ability to form letters is assumed at all levels. **Orthographic competence** is assessed in the interactive writing tasks and is captured in the assessment criteria of "**Accuracy**"
After reading the scales for Range and Accuracy in Table 4.4, indicate and justify at which level(s) of the scale the examination should be situated.	**Level B2** ***Justification*** ***(incl. reference to documentation)*** The assessment standards, exam tasks, input texts and assessment criteria for the International ESOL examination were drawn up using the CEF. Shows a relatively high degree of grammatical control Does not make mistakes which lead to misunderstanding. For more evidence see: International ESOL Handbook Item Writer & Editor Guide Marking Guide

Form A20a: Aspects of Linguistic Competence in Interaction

SOCIO-LINGUISTIC COMPETENCE	*IESOL*
What are the socio-linguistic competences that the test takers are expected to be able to handle: linguistic markers politeness conventions, register, adequacy, dialect/accent, etc.?	For Writing Skills this socio-linguistic element is captured in the candidate's ability to complete the task. Its chief aspects include: candidate awareness of social conventions, register and politeness. This sociolinguistic element is captured in the assessment criteria for task completion i.e. "Global" assessment criteria, as well as the assessment criteria for "Range" and "Organisation". Candidates are required to produce texts with varying degree of formality/ register using the correct conventions and appropriate linguistic differences Also required to address finer issues of sociolinguistic competence i.e. "adjusting register to suit purpose & readership"
After reading the scale for Socio-linguistic Competence in Table 4.4, indicate and justify at which level(s) of the scale the examination should be situated.	**Level B2** ***Justification (incl. reference to documentation)*** The assessment standards, exam tasks, assessment criteria for the International ESOL examination were drawn up from the CEF. Can with some effort keep up with and contribute to group discussions even when speech is fast and colloquial. Can sustain relationships with native speakers without intentionally amusing or irritating them or requiring them to behave other than they would with a native speaker. Can express him or herself appropriately in situations and avoid crass errors of formulation. For more evidence see: International ESOL Handbook Item writer's and editor's checklist Assessment Criteria in the Marking Guide

Form A20b: Aspects of Socio-Linguistic Competence in Interaction

PRAGMATIC COMPETENCE	
What are the pragmatic competences that the test takers are expected to be able to handle: discourse competences, functional competences?	**Pragmatic Competences:** **Discourse Competence** Candidates are expected to display the discourse competence of **coherence and cohesion and thematic development**. Thematic Development is captured in the candidate's ability to complete the task ("Global" Assessment Criteria) and Coherence & Cohesion is captured in the assessment criteria of "Organisation" **Functional Competences**- the candidates are expected to handle the "Function lists" given for each level in the International ESOL handbook. The main headings for each level are as follows: Personal environment, Expressing thoughts and feelings, making things happen, social contact and expressing attitudes
After reading the scale for Fluency in Table 4.4, indicate and justify at which level(s) of the scale the examination should be situated.	**Level B2** ***Justification (incl. reference to documentation)*** The assessment standards, exam tasks, assessment criteria for the International ESOL examination were drawn up from the CEF. Can use a limited number of cohesive devices to link his/her utterances into clear coherent discourse, though there maybe some jumpiness in along contribution. For more evidence see: International ESOL Handbook Item writer's and editor's checklist Assessment Criteria in the Marking Guide

Form A20c: Aspects of Pragmatic Competence in Interaction

STRATEGIC COMPETENCE	
What are the interaction strategies that the test takers are expected to be able to handle?	Interaction strategies are tested indirectly in the candidate's ability to complete the interactive tasks in the writing section of the examination.
After reading the scale for Interaction in Table 4.4, indicate and justify at which level(s) of the scale the examination should be situated.	**Level B2** ***Justification*** ***(incl. reference to documentation)*** The assessment standards, exam tasks, input texts and assessment criteria for the International ESOL examination were drawn up from the CEF. Can initiate discourse, take his/her turn when appropriate and end conversation when he/she needs to, although may not always do this elegantly. Can help the discussion along on familiar ground confirming comprehension, inviting others in and so on. For more evidence see: International ESOL Handbook Item writer's and editor's checklist Assessment Criteria in the Marking Guide

Form A20d: Aspects of Strategic Competence in Interaction

LINGUISTIC COMPETENCE	
What is the range of lexical and grammatical competence that the test takers are expected to be able to handle?	For a detailed listing of the grammatical competence expected at each level please see the International ESOL Qualification Handbook pp 79-86 Lexical competence in Written Production is captured in the assessment criteria of "**Range**"
What is the range of phonological and orthographic competence that the test takers are expected to be able to handle?	Familiarity with the Roman alphabet and ability to form letters is assumed at all levels. **Orthographic competence** is assessed in the writing tasks and is captured in the assessment criteria of "**Accuracy**"
After reading the scales for Range and Accuracy in Table 4.4, indicate and justify at which level(s) of the scale the examination should be situated.	**Level B2** ***Justification*** ***(incl. reference to documentation)*** The assessment standards, exam tasks, input texts and assessment criteria for the International ESOL examination were drawn up using the CEF. Shows a relatively high degree of grammatical control Does not make mistakes which lead to misunderstanding. For more evidence see: International ESOL Handbook Item Writer & Editor Guide Marking Guide

Form A21a: Aspects of Linguistic Competence in Production

SOCIO-LINGUISTIC COMPETENCE	
What are the socio-linguistic competences that the test takers are expected to be able to handle: linguistic markers politeness conventions, register, adequacy, dialect/accent, etc?	*For Writing Skills this socio-linguistic element is captured in the candidate's ability to complete the task. Its chief aspects include: candidate awareness of social conventions, register and politeness. This sociolinguistic element is captured in the assessment criteria for task completion i.e. "Global" assessment criteria, as well as the assessment criteria for "Range" and "Organisation".* candidates are required to produce texts with varying degree of formality/ register using the correct conventions and appropriate linguistic differences Also required to address finer issues of sociolinguistic competence i.e. "adjusting register to suit purpose & readership"
After reading the scale for Socio-linguistic Competence in Table 4.5, indicate and justify at which level(s) of the scale the examination should be situated.	**Level B2** ***Justification*** ***(incl. reference to documentation)*** The assessment standards, exam tasks, assessment criteria for the International ESOL examination were drawn up from the CEF. Can with some effort keep up with and contribute to group discussions even when speech is fast and colloquial. Can sustain relationships with native speakers without intentionally amusing or irritating them or requiring them to behave other than they would with a native speaker. Can express him or herself appropriately in situations and avoid crass errors of formulation. For more evidence see: International ESOL Handbook Item writer's and editor's checklist Assessment Criteria in the Marking Guide

Form A21b: Aspects of Socio-Linguistic Competence in Production

PRAGMATIC COMPETENCE	
What are the pragmatic competences that the test takers are expected to be able to handle: discourse competences, functional competences?	**Pragmatic Competences:** **Discourse Competence** Candidates are expected to display the discourse competence of **coherence and cohesion and thematic development**. Thematic Development is captured in the candidate's ability to complete the task ("Global" Assessment Criteria) and Coherence & Cohesion is captured in the assessment criteria of "Organisation" **Functional Competences** Candidates are expected to handle the "Function lists" given for each level in the International ESOL handbook. The main headings for Communicator can be found on pages 91-93.
After reading the scale for Pragmatic Competence in Table 4.4, indicate and justify at which level(s) of the scale the examination should be situated.	**Level B2** ***Justification*** ***(incl. reference to documentation)*** The assessment standards, exam tasks, assessment criteria for the International ESOL examination were drawn up from the CEF. **Communicator Level = CEF Description for Coherence and Cohesion B2:** Can use a limited number of cohesive devices to link his/her utterances into clear coherent discourse, though there maybe some jumpiness in along contribution. For more evidence see: International ESOL Handbook Item writer's and editor's checklist Assessment Criteria in the Marking Guide

Form A21c: Aspects of Pragmatic Competence in Production

STRATEGIC COMPETENCE	
What are the production strategies that the test takers are expected to be able to handle?	Production strategies are tested indirectly by candidate's ability to complete writing tasks. They are expected to include: **Planning** • Language & content review; • Locating resources; • Considering audience; • Task adjustment; • Message adjustment. **Execution** • Building on previous knowledge; • On-line language review. **Evaluation** • Monitoring written performance. **Repair** • Editing (while and post writing) • Self-correction.
After reading the scale for Strategic Competence in Table 4.4, indicate and justify at which level(s) of the scale the examination should be situated.	**Level B2** ***Justification*** ***(incl. reference to documentation)*** The assessment standards, exam tasks, assessment criteria for the International ESOL examination were drawn up from the CEF. Can use circumlocution and paraphrase to cover gaps in vocabulary and structure. Can consciously monitor own linguistic production For more evidence see: International ESOL Handbook Item writer's and editor's checklist Assessment Criteria in the Marking Guide

Form A21: Aspects of Strategic Competence in Production

C2.2	Mastery (first class pass)	Mastery (first class pass)	Mastery (first class pass)	Mastery (first class pass)
C2	Mastery (pass)	Mastery (pass)	Mastery (pass)	Mastery (pass)
C1.2	Expert (first class pass)	Expert (first class pass)	Expert (first class pass)	Expert (first class pass)
C1	Expert (pass)	Expert (pass)	Expert (pass)	Expert (pass)
B2.2	Communicator (first class pass)	Communicator (first class pass)	Communicator (first class pass)	Communicator (first class pass)
B2	Communicator (pass)	Communicator (pass)	Communicator (pass)	Communicator (pass)
B1.2	Achiever (first class pass)	Achiever (first class pass)	Achiever (first class pass)	Achiever (first class pass)
B1	Achiever (pass)	Achiever (pass)	Achiever (pass)	Achiever (pass)
A2.2	Access (first class pass)	Access (first class pass)	Access (first class pass)	Access (first class pass)
A2	Access (pass)	Access (pass)	Access (pass)	Access (pass)
A1.2	Preliminary (first class pass)	Preliminary (first class pass)	Preliminary (first class pass)	Preliminary (first class pass)
A1	Preliminary (pass)	Preliminary (pass)	Preliminary (pass)	Preliminary (pass)
Overall	**IESOL Overall**	**IESOL Listening subtest**	**IESOL Reading subtest**	**IESOL Writing subtest**

Form A23: Graphic Profiles of the Relationship of the Examination to CEF Levels

Appendix 3 Extracts from FACETS output (Preliminary Expert Panel)

Writing Results

```
C&G Writing Text  06-19-2007 18:35:55
Table 6.0  All Facet Vertical "Rulers".

Vertical = (1A,2A,3A,4A) Yardstick (columns,lines,low,high)= 0,10,-1,1
-----------------------------------------------
|Measr|+Task           |-Raters |-Output|S.1  |
-----------------------------------------------
+   1 +                +        +       +(5)  +
|     |                |        |       |     |
|     |                |        |       | 4   |
|     |                |        |       |     |
|     |                |        |       |     |
|     |                |        |       |     |
|     |                |        |       |     |
|     |                | OD     |       |     |
|     |                |        | 2     |     |
|     |                | OS     |       |     |
*   0 *                * IR     * 1     * --- *
|     |                |        |       |     |
|     |                | IS  OC | 3     |     |
|     | TASK 4         |        |       |     |
|     |                |        |       |     |
|     |                |        |       |     |
|     |                |        |       |     |
|     |                |        |       |     |
|     | TASK 1  TASK 3 |        |       | 3   |
|     | TASK 2         |        |       |     |
+  -1 +                +        +       +(2)  +
-----------------------------------------------
|Measr|+Task           |-Raters |-Output|S.1  |
-----------------------------------------------


C&G Writing Text  06-19-2007 18:35:55
Table 7.1.1  Task Measurement Report  (arranged by 1N).

-------------------------------------------------------------------------------------------
| Obsvd   Obsvd  Obsvd  Fair-M|        Model | Infit      Outfit   |
| Score   Count Average Avrage|Measure  S.E. |MnSq ZStd  MnSq ZStd | N Task
-------------------------------------------------------------------------------------------
|     45     15     3.0   3.00|   -.82   .34 | 1.0   0    1.0   0  | 1 TASK 1
|     44     15     2.9   2.93|   -.94   .34 |  .9   0     .9   0  | 2 TASK 2
|     45     15     3.0   3.00|   -.82   .34 | 1.3   0    1.3   0  | 3 TASK 3
|     50     15     3.3   3.33|   -.28   .33 |  .8   0     .8   0  | 4 TASK 4
-------------------------------------------------------------------------------------------
|    46.0    15.0   3.1   3.07|   -.71   .34 | 1.0    .0  1.0    .0| Mean (Count: 4)
|     2.3      .0    .2    .16|    .26   .01 |  .2    .6   .2    .6| S.D.
-------------------------------------------------------------------------------------------
RMSE (Model)  .34 Adj S.D.   .00  Separation   .00  Reliability  .00
Fixed (all same) chi-square: 2.4  d.f.: 3  significance: .49
-------------------------------------------------------------------------------------------
```

C&G Writing Text 06-19-2007 18:35:55
Table 7.2.1 Raters Measurement Report (arranged by fN).

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	N	Raters
38	12	3.2	3.17	-.17	.37	1.2	0	1.2	0	4	IS
38	12	3.2	3.17	-.17	.37	1.2	0	1.2	0	3	OC
37	12	3.1	3.08	-.03	.38	.9	0	.9	0	5	IR
36	12	3.0	3.00	.11	.38	.9	0	.9	0	1	OS
35	12	2.9	2.92	.26	.39	.8	0	.8	0	2	OD
36.8	12.0	3.1	3.07	.00	.38	1.0	-.1	1.0	.0	Mean (Count: 5)	
1.2	.0	.1	.10	.17	.01	.2	.5	.1	.4	S.D.	

RMSE (Model) .38 Adj S.D. .00 Separation .00 Reliability .00
Fixed (all same) chi-square: .9 d.f.: 4 significance: .92

Reading Results

```
C&G Reading Text  06-19-2007 18:35:32
Table 6.0  All Facet Vertical "Rulers".

Vertical = (1A,2A,3A) Yardstick (columns,lines,low,high)= 0,5,-6,3
-------------------------------------------------
|Measr|+Task                      |-Raters|S.1  |
-------------------------------------------------
+   3 +                           +       +(5)  +
|     |                           |       |     |
|     | TASK 3                    |       | 4   |
|     |                           |       |     |
|     |                           | OD    |     |
+   2 +                           +       +     +
|     | TASK 8                    |       |     |
|     |                           | OC    |     |
|     |                           |       |     |
|     |                           |       |     |
+   1 + TASK 9                    +       +     +
|     |                           |       |     |
|     |                           | IR    | --- |
|     |                           |       |     |
|     |                           |       |     |
*   0 * TASK 1   TASK 10  TASK 7  *       *     *
|     |                           |       |     |
|     |                           | OS    |     |
|     |                           |       |     |
|     |                           |       |     |
+  -1 +                           +       +     +
|     | TASK 4                    |       |     |
|     |                           |       |     |
|     |                           |       |     |
|     |                           |       |     |
+  -2 +                           +       +     +
|     |                           |       | 3   |
|     |                           |       |     |
|     |                           |       |     |
|     |                           |       |     |
+  -3 +                           +       +     +
|     |                           |       |     |
|     |                           |       |     |
|     |                           |       |     |
|     |                           |       |     |
+  -4 +                           + IS    +     +
|     |                           |       |     |
|     | TASK 2   TASK 5           |       |     |
|     |                           |       |     |
|     |                           |       |     |
+  -5 +                           +       +     +
|     |                           |       | --- |
|     |                           |       |     |
|     | TASK 6                    |       |     |
|     |                           |       |     |
+  -6 +                           +       +(2)  +
-------------------------------------------------
|Measr|+Task                      |-Raters|S.1  |
-------------------------------------------------
```

C&G Reading Text 06-19-2007 18:35:32
Table 7.1.1 Task Measurement Report (arranged by 1N).

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Nu	Task
17	5	3.4	3.31	-.05	1.04	.6	0	.4	0	1	TASK 1
13	5	2.6	2.68	-4.37	1.00	1.3	0	1.3	0	2	TASK 2
20	5	4.0	4.03	2.66	.94	.6	0	.5	0	3	TASK 3
16	5	3.2	3.10	-1.24	1.15	.1	-1	.1	-1	4	TASK 4
13	5	2.6	2.68	-4.37	1.00	1.3	0	1.3	0	5	TASK 5
12	5	2.4	2.41	-5.51	1.16	1.0	0	.7	0	6	TASK 6
17	5	3.4	3.31	-.05	1.04	.6	0	.4	0	7	TASK 7
19	5	3.8	3.81	1.80	.93	1.1	0	1.1	0	8	TASK 8
18	5	3.6	3.57	.92	.96	.4	-1	.4	-1	9	TASK 9
17	5	3.4	3.31	-.05	1.04	2.0	1	2.7	1	10	TASK 10
16.2	5.0	3.2	3.22	-1.02	1.02	.9	-.3	.9	-.3	Mean (Count: 10)	
2.6	.0	.5	.49	2.66	.08	.5	.9	.7	.9	S.D.	

RMSE (Model) 1.03 Adj S.D. 2.46 Separation 2.39 Reliability .85
Fixed (all same) chi-square: 68.7 d.f.: 9 significance: .00

C&G Reading Text 06-19-2007 18:35:32
Table 7.2.1 Raters Measurement Report (arranged by fN).

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	N	Raters
28	10	2.8	2.89	2.20	.85	1.4	0	1.5	0	2	OD
33	10	3.3	3.18	-.34	.65	1.1	0	1.2	0	1	OS
31	10	3.1	3.06	.53	.68	.8	0	.7	0	5	IR
29	10	2.9	2.96	1.56	.77	.9	0	.7	0	3	OC
41	10	4.1	4.09	-3.94	.69	.5	-1	.4	-1	4	IS
32.4	10.0	3.2	3.24	.00	.73	.9	-.3	.9	-.3	Mean (Count: 5)	
4.6	.0	.5	.44	2.15	.07	.3	.7	.4	.8	S.D.	

RMSE (Model) .73 Adj S.D. 2.02 Separation 2.77 Reliability .88
Fixed (all same) chi-square: 43.6 d.f.: 4 significance: .00

Listening Results

```
C&G Listening Text  06-19-2007 18:27:23
Table 6.0  All Facet Vertical "Rulers".

Vertical = (1A,2A,3A) Yardstick (columns,lines,low,high)= 0,7,-5,2
-------------------------------------------------------------
|Measr|+Task                                |-Raters |S.1  |
-------------------------------------------------------------
+   2 +                                    +        +(6)  +
|     |                                    |        |     |
|     |                                    |        |     |
|     |                                    |        | --- |
|     |                                    |        |     |
|     |                                    |        |     |
|     |                                    | OC  OD |     |
+   1 +                                    +        +     +
|     |                                    |        |     |
|     |                                    | OS     | 4   |
|     |                                    |        |     |
|     |                                    |        |     |
|     |                                    |        |     |
|     | TASK 3                             |        |     |
*   0 *                                    *        *     *
|     |                                    |        |     |
|     |                                    |        |     |
|     | TASK 2   TASK 5   TASK 6   TASK 7  |        | --- |
|     |                                    |        |     |
|     |                                    |        |     |
|     |                                    |        |     |
+  -1 +                                    +        +     +
|     |                                    | IR     |     |
|     |                                    |        |     |
|     |                                    |        |     |
|     |                                    |        |     |
|     |                                    | IS     |     |
|     |                                    |        |     |
+  -2 +                                    +        +     +
|     | TASK 8   TASK 9                    |        | 3   |
|     |                                    |        |     |
|     |                                    |        |     |
|     |                                    |        |     |
|     |                                    |        |     |
|     |                                    |        |     |
+  -3 +                                    +        +     +
|     |                                    |        |     |
|     |                                    |        |     |
|     |                                    |        |     |
|     | TASK 1                             |        |     |
|     |                                    |        |     |
|     |                                    |        |     |
+  -4 +                                    +        + --- +
|     |                                    |        |     |
|     |                                    |        |     |
|     |                                    |        |     |
|     | TASK 10  TASK 4                    |        |     |
|     |                                    |        |     |
|     |                                    |        |     |
+  -5 +                                    +        +(2)  +
-------------------------------------------------------------
|Measr|+Task                                |-Raters |S.1  |
-------------------------------------------------------------
```

C&G Listening Text 06-19-2007 18:27:23
Table 7.1.1 Task Measurement Report (arranged by 1N).

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Nu	Task
13	5	2.6	2.61	-3.60	.90	.6	0	.6	0	1	TASK 1
18	5	3.6	3.52	-.37	.70	1.3	0	1.7	0	2	TASK 2
19	5	3.8	3.69	.07	.65	.6	0	.7	0	3	TASK 3
12	5	2.4	2.36	-4.51	1.01	1.6	0	1.8	0	4	TASK 4
18	5	3.6	3.52	-.37	.70	1.8	0	1.2	0	5	TASK 5
18	5	3.6	3.52	-.37	.70	.3	-1	.2	-1	6	TASK 6
18	5	3.6	3.52	-.37	.70	.3	-1	.4	-1	7	TASK 7
15	5	3.0	2.99	-2.15	.82	.6	0	.6	0	8	TASK 8
15	5	3.0	2.99	-2.15	.82	.9	0	.9	0	9	TASK 9
12	5	2.4	2.36	-4.51	1.01	1.5	0	1.4	0	10	TASK 10
15.8	5.0	3.2	3.11	-1.83	.80	1.0	-.2	1.0	-.2	Mean (Count: 10)	
2.6	.0	.5	.49	1.73	.13	.5	.8	.5	.8	S.D.	

RMSE (Model) .81 Adj S.D. 1.53 Separation 1.89 Reliability .78
Fixed (all same) chi-square: 40.5 d.f.: 9 significance: .00

C&G Listening Text 06-19-2007 18:27:23
Table 7.2.1 Raters Measurement Report (arranged by fN).

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	N	Raters
28	10	2.8	2.80	1.08	.64	1.2	0	1.4	0	3	OC
29	10	2.9	2.90	.70	.61	1.3	0	1.3	0	1	OS
38	10	3.8	3.63	-1.76	.44	1.2	0	1.3	0	4	IS
35	10	3.5	3.40	-1.11	.49	.5	-1	.5	-1	5	IR
28	10	2.8	2.80	1.08	.64	.4	-1	.3	-1	2	OD
31.6	10.0	3.2	3.10	.00	.56	.9	-.3	1.0	-.2	Mean (Count: 5)	
4.1	.0	.4	.35	1.19	.08	.4	1.0	.5	1.1	S.D.	

RMSE (Model) .57 Adj S.D. 1.05 Separation 1.85 Reliability .77
Fixed (all same) chi-square: 25.5 d.f.: 4 significance: .00

Appendix 4 Self Assessment 'Can DO' Instrument

SELF ASSESSMENT QUESTIONNAIRE

Please show your level of agreement with these statements
(1 = I really do not agree; 5 = I fully agree)

		NOT AGREE 1 2 3 4 5 AGREE
READING	I can deal with routine letters.	1 2 3 4 5
	I can understand at least the general meaning of more complex articles.	1 2 3 4 5
	I can understand most short reports provided I have enough time.	1 2 3 4 5
	I can follow simple instructions given on packaging (e.g. cooking instructions on a packet of pasta).	1 2 3 4 5
	I can understand a factual article or report in a newspaper or magazine.	1 2 3 4 5
	I can understand instructions for things that are outside my job area.	1 2 3 4 5
	I can understand the general meaning of a theoretical article within own work area of study.	1 2 3 4 5
	I can understand most reports but only if I have a lot of time.	1 2 3 4 5
	I can understand letters even when they contain non-standard language.	1 2 3 4 5
	I can understand operating instructions on electrical appliances, e.g. an iPod.	1 2 3 4 5
	I can understand information given in guide books.	1 2 3 4 5
	I can understand complex opinions and arguments as expressed in serious newspapers.	1 2 3 4 5
	I can understand opinions where these are simply expressed.	1 2 3 4 5
	I can deal with any advertisement for a flat and understand most of the abbreviations and terms used.	1 2 3 4 5
	I can read newspapers or the internet for information quickly with no problems.	1 2 3 4 5

		NOT AGREE 1 2 3 4 5 AGREE
LISTENING	I have no problem understanding casual conversations about complicated topics even for a long time.	1 2 3 4 5
	I can understand almost all of what people say on the telephone.	1 2 3 4 5
	I can follow most of what is said in a lecture, presentation or demonstration.	1 2 3 4 5
	I can take fairly complex messages on the phone, provided the caller speaks slowly and carefully.	1 2 3 4 5
	I can understand most explanations of what is on a menu, but I need a dictionary for very specific words about food.	1 2 3 4 5
	I can understand in detail an argument in a discussion programme on radio or TV.	1 2 3 4 5
	I can understand simple answers to questions I ask when I am in a restaurant or shop.	1 2 3 4 5
	I can understand a casual conversation for a fairly long period of time if the topic is easy.	1 2 3 4 5
	I can follow what is said at a meeting, though I may need to ask the speaker to explain some parts for me.	1 2 3 4 5
	I can understand most of what is said on most guided tours.	1 2 3 4 5
	I can understand the main topic of a news programme on TV when there are photos or videos.	1 2 3 4 5
	I can tell when the lecturer makes an important point from when he/she is just giving extra information.	1 2 3 4 5
	I can understand most of what takes place at a meeting.	1 2 3 4 5
	I can follow the development of a discussion during a seminar.	1 2 3 4 5
	I can understand most of what is said in a TV or radio programme, or in a film.	1 2 3 4 5

Appendix 5 Sample Communicator Paper (CEFR B2)

International ESOL
English for Speakers of Other Languages
Communicator Level – B2
Practice Paper 1

This paper must be returned with the candidate's work, otherwise the entry will be void and no result will be issued.

8984-74-074
(EL-IESOL 4)

City & Guilds new 2009 CEFR aligned Practice Paper

Candidate's name (block letters please)

Centre no **Date**

Time allowed: **2 hours and 30 minutes**

- Listening

- Reading

- Writing

Instructions to Candidates

- Answer all the questions.

- All your answers must be written in **ink** not pencil.

For examiner's use only

Parts	L1	L2	L3	L4	Total	R1	R2	R3	R4	Total	W1	W2
Candidate's score												
RESULTS: LISTENING						READING					WRITING	
OVERALL RESULT:												

Listening Part 1

You will hear eight short unfinished conversations. Choose the best reply to continue the conversation. Put a circle round the letter of the best reply. You will hear the conversations once only. First, look at the example.

Example:

Speaker 1: *Are you sure this one will fit into the room?*
Speaker 2: *It's no bigger than the one we have now.*
Speaker 1: *You really should measure it.*
Speaker 2: .....

a) Why are you so surprised?
b) You worry too much.
c) I'll change it after I finish this one.
d) I have it right here.

1. a) You'll need a doctor's note.
   b) You could do much better.
   c) Don't be afraid of it.
   d) That's not reasonable.

2. a) We'll need to organise the event.
   b) They won't all fit in at the same time.
   c) That's the best idea so far today.
   d) We could try giving better directions.

3. a) I'm not keen on having another.
   b) But I have to go to work now.
   c) I'll make time for you.
   d) But I want to order it now.

4. a) It seems an impossible job.
   b) I know. I didn't believe them either.
   c) You could see how it was done.
   d) I know. I really felt I was there.

5. a) You're bound to add more.
   b) OK, but that's it. I'm off now.
   c) I'll get quite a bit, then.
   d) See you next week, then.

6. a) I'll have to write it down.
   b) Sorry, I don't understand your problem.
   c) Perhaps we'll find it somewhere.
   d) Thanks. I get what you mean now.

7. a) Sorry, she's out all day today.
   b) Could you tell me who you need to see?
   c) Could you hold, please, and I'll check?
   d) Sorry, but it's an expensive call.

8. a) We've not met for ages.
   b) It's always a pleasure.
   c) I'm a lot older too.
   d) I've heard all about you.

(Total: 8 marks)

Listening Part 2

You will hear three conversations. Listen to the conversations and answer the questions below. Put a circle round the letter of the correct answer. You will hear each conversation once only. Look at the questions for Conversation One.

Conversation 1

1.1 The man and woman are

a) buying a house.
b) planning a garden.
c) looking for a new hobby.
d) discussing cookery.

1.2 The man is

a) excited.
b) frightened.
c) annoyed.
d) surprised.

Conversation 2

2.1 The speakers are talking about

a) a murder.
b) a mugging.
c) shoplifting.
d) a burglary.

2.2 The man and woman

a) work together.
b) live together.
c) are neighbours.
d) are teachers.

Conversation 3

3.1 The man is

a) in a chemist's shop.
b) in a shoe shop.
c) at a doctor's surgery.
d) in a clothes shop.

3.2 The woman is

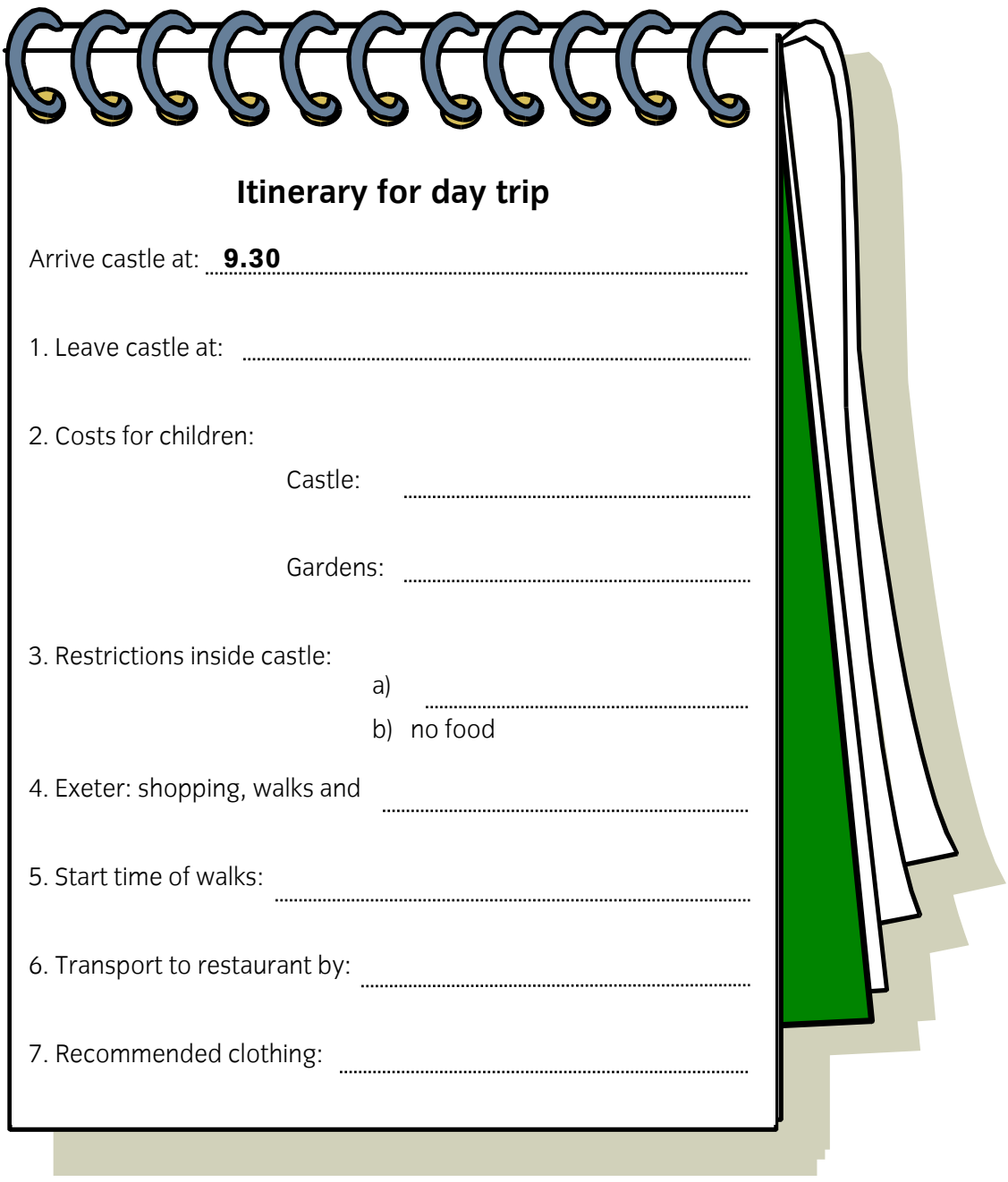
a) offering congratulations.
b) giving advice.
c) paying a compliment.
d) giving praise.

(Total: 6 marks)

Listening Part 3

*Listen to the message about a day trip. Make **short** notes about the message. First, look at the notes.*

The first one is done for you. You will hear the message once only.



Itinerary for day trip

Arrive castle at: **9.30**

1. Leave castle at: .....

2. Costs for children:

   Castle: .....

   Gardens: .....

3. Restrictions inside castle:

   a) .....

   b) no food

4. Exeter: shopping, walks and .....

5. Start time of walks: .....

6. Transport to restaurant by: .....

7. Recommended clothing: .....

(Total: 8 marks)

Listening Part 4

Listen to the conversation and answer the questions. Put a circle round the letter of the correct answer. First, look at the questions. The first one is done for you. You will hear the conversation twice.

Example:

John will be having dinner

a) at home with his parents.
b) at a friend's house.
c) at the cinema.
d) at work.

1. What would John's father like his son to take more seriously?

   a) Football.
   b) Family life.
   c) Education.
   d) Food.

2. John and his parents live in

   a) an urban area.
   b) a suburban area.
   c) a remote area.
   d) a rural area.

3. John's father initially thinks that buying his son a car is

   a) a terrible idea.
   b) an impossibility.
   c) an absurd idea.
   d) a waste of time.

4. John's mother considers her son to be

   a) energetic.
   b) studious.
   c) lazy.
   d) sociable.

5. John's mother changed her mind about the car because

   a) John can persuade her very easily.
   b) John gave good reasons to have one.
   c) She thinks in the same way.
   d) She's always a good listener.

6. Why doesn't John have a part-time job?

a) He's always doing something.
b) He has to study all the time.
c) He doesn't have transport.
d) He doesn't need the money.

7. John's mother is in favour of

a) buying a cheap car.
b) lending John her car.
c) buying an economical car.
d) giving John money to buy a car.

8. What must John do before he gets a car?

a) Pass his final school exams.
b) Learn about car maintenance.
c) Get accepted at university.
d) Pass his driving test and get a job.

(Total: 8 marks)

Reading Part 1
Read the text and complete the tasks that follow. Choose a, b, c or d. Put a circle round the most appropriate answer. An example is done for you.

Lottery winners who lose their millions

For a lot of people, winning the lottery is a dream come true. But for many, the reality is more like a nightmare.

Evelyn Adams won $5.4 on the New Jersey lottery in 1986. Today the money is all gone and Adams lives in a trailer. 'Everybody wanted my money. I never learned to say 'No.' I wish I had the chance to do it all over again. I'd be much smarter about it now. I was a big-time gambler,' admits Adams. 'I made mistakes, some I regret, some I don't. I can't go back now so I just go forward, one step at a time.'

William 'Bud' Post won $16.2 million in the Pennsylvania lottery in 1988. 'I wish it never happened. It was totally a nightmare,' says Post. A former girlfriend successfully sued him for a share of his winnings, a brother was arrested for hiring a hit man to kill him, hoping to inherit a share of the winnings. Other siblings persuaded him to invest in a car showroom and a restaurant, both of which failed through his mismanagement and further strained family relationships. Post now lives quietly on $450 a month, having lost virtually all his money.

Ken Proxmire was a machinist when he won $1 million in the Michigan lottery. He moved to California and went into the car business with his brothers. Within five years, he had filed for bankruptcy. 'He was just a poor boy who got lucky and wanted to take care of everybody,' explains Ken's son Rick. It was a hell of a good ride for three or four years, but now he lives more simply working as a machinist,' says his son.

These sad-but-true tales are not uncommon, says Susan Bradley, a certified financial planner. 'There is a widely held belief that money solves problems. But people soon learn that money can cause as many problems as it solves,' she says.

Bradley recommends taking time out from making any financial decisions. 'It's a time to think things through, sort things out and only then to seek an advisory team to help make those important financial choices,' she says. 'You really don't want to buy a new house before taking the time to think about what the consequences are. People don't realise how much it costs to live in a big house – decorators, furniture, taxes, insurance, even utility costs are greater. People need a reality check before they sign the contract.'

Example:

For many lottery winners the dream

a) can become reality.
b) is not always a good one.
c) is better than they imagined.
d) can remain just a dream.

1. For Evelyn, winning the lottery

a) has taught her a lot about life.
b) was the best thing to happen to her.
c) brought her closer to her family.
d) is something she regrets.

2. William Post's ex-girlfriend

a) was taken to court by him.
b) bought the winning lottery ticket.
c) stole some of his money.
d) took legal action against him.

3. Post lost a lot of his money because

a) he wasn't a good businessman.
b) his brothers and sisters tricked him.
c) he got on badly with his family.
d) he gave too much of it away.

4. According to Ken Proxmire's son, his father was

a) not used to having money.
b) lucky throughout his life.
c) too concerned about others.
d) rich for about five years.

5. Susan Bradley thinks lottery winners should begin by

a) developing a financial partnership.
b) starting financial planning.
c) not asking experts to help them.
d) thinking instead of spending.

6. In summary, the article says that, if you win a lot of money,

a) don't take anyone else's advice.
b) don't assume it will make you happy.
c) put some of it away in a bank.
d) treat family members with suspicion.

(Total: 6 marks)

Reading Part 2

Read the text and fill the gaps with sentences A - H. Write the letter of the missing sentence in the box in the correct gap. There are two extra sentences you will not need.

England's disappearing coastline

Explore north-east Norfolk before the tide comes in. It's from the top of Horsey Mill that the problem becomes apparent. [ ] **1.** The artificially raised dunes along the coast provide little protection for land lower than the waves that rage at its door.

Depending on whom you believe, this 25-square-mile triangle of north-east Norfolk will be reclaimed by the North Sea in 20-50 years' time. [ ] **2.** Even worse, some experts believe that the sea could come in at any time and flood it.

Up the road is Waxham Great Barn. [ ] **3.** It is possibly the longest thatched barn in Britain, sitting in the valley surrounded by ancient woodland, home to a colony of Natterer's bats. House martins swoop overhead as I sip. This has to be the most magnificent setting for a teashop anywhere in the land. [ ] **4.**

In this beautiful spot, peace and quiet reign. But the sea is ever present. [ ] **5.** Earlier I went for a stroll along a wonderfully deserted sandy beach. The dunes are so precarious, the paths along them have been closed. [ ] **6.**

A Here, the green of the fields lies as flat as the blue of the meres: you could iron a shirt on it.

B I can hear the long, withdrawing roar of waves a quarter of a mile away from my campsite.

C It was built with the stone of three old monasteries to create something resembling a cathedral.

D This made me realise just how much damage has already been done.

E Maintaining the nine miles of defenses after then is apparently unsustainable.

F Nevertheless, I was unprepared to witness the full extent of the sea's destructive power.

G If this isn't what the English countryside should be, I don't know what is.

H What appears at first sight to be a little piece of paradise is consequently a disaster waiting to happen.

(Total: 6 marks)

Reading Part 3
Read the four texts below. There are ten questions about the texts. Decide which text A, B, C or D tells you the answer to the question. The first one is done for you.

A

The fascination with medieval Islamic architecture that pervades paintings such as John Frederick Lewis's The Bezestein Bazaar of El Khan Khalil, Cairo (1872) makes for superb portrayals of some of the world's great urban spaces. His watercolours are incredibly fine notations of the stucco-work and the tiles, lattices and niches that make Islamic architecture in many ways the most beautiful ever created. It is hard to discern any underlying imperial disdain. None of these painters is a great artist, and yet the exhibition is full of great art.

B

£10 (£9 Senior Citizen, £8 Student/Job Seeker/Child 12-18 yrs/Disabled concessions)
Free for Tate Members
Book online with Tate or call 020 7887 8888

Tickets for special exhibitions can be bought at Tate Britain or Tate Modern seven days a week from 10.00 to 17.00, with late opening until 21.00 at Tate Modern on Friday and Saturday.

There is no booking fee when you buy tickets in person at the galleries. We do however encourage you to purchase tickets in advance online.

C

I've bought our tickets for the exhibition so that we don't have to queue this evening. I'm good, aren't I?

Anyway, I'll meet you at the Gallery restaurant, near Tate Britain, at 6.30 pm. That way we can have dinner before we catch the late showing which is open until 9pm tonight. The restaurant is meant to be really good! I think an hour and a half should give us enough time to see the art work, don't you? See you later.

D

Thank you for your query about future exhibitions on analogous themes to the 'Orient' one. I'm afraid that there are none planned at present. However, I added you to our mailing list, so you will be informed of all forthcoming events.

We greatly welcome feedback from visitors, and wondered if you wish to contribute to our monthly newsletter. You might be interested to know that there are special concessions for 'Friends of the Tate' who assist us in this way.

Which text:

1.	provides information about opening times?	**B**
2.	invites the public's opinions?	
3.	refers to more than one gallery?	
4.	describes the subject matter of the works?	
5.	indicates where you can see the exhibition?	

Which text gives you the answers to the following questions?

6.	What is the best way to ensure entry to the exhibition?	
7.	What's the best way to learn about future exhibitions?	
8.	Which materials were used in the paintings?	
9.	How long does it take to see the exhibition?	
10.	How can you show your support for the gallery?	

(Total: 9 marks)

Reading Part 4
Read the text and answer the questions. ***Write a maximum of five words for each answer****. An example is done for you.*

How much faster can humans run?

When Usain Bolt, a 21-year-old Jamaican, smashed the world 100-metres record in the Beijing Olympic Final in 2008, it was the 18th time the record had been legally broken since an American called Don Lippincott ran 10.6 seconds in 1912, and the 8th new 100 metres record set since 1991. The 10-second barrier was broken in 1968, the 9.90 barrier in 1991, and the 9.80 barrier in 1999. Now, with the record standing at 9.69, the 9.70 barrier has also been broken.

The best sprinters are running, albeit briefly, at about 26-27mph. The title of 'fastest man in the world' is traditionally held by the 100-metre world record holder, but one scientific form of reckoning bestows that title on the former 200-metre runner Michael Johnson, whose performance in setting the world record of 19.32 sec at the 1996 Olympics produced an average speed of 23.15mph (compared with Bolt's 23.02mph in his record-breaking run). In terms of peak speed, Canada's Donovan Bailey is credited with the record, hitting 27.07mph in winning the 100m title at the 1996 Olympics in a then world record of 9.84 sec.

The days when 100-metre runners used to knock a tenth of a second off the world record – as Jesse Owens did in running 10.2 sec in 1936 – are long gone. The record has been creeping down in hundredths of second since Jim Hines became the first man to break 10 seconds in 1968, winning the Olympic title in 9.95 sec.

Ben Johnson was infamously stripped of his 1988 Olympic 100-metre title – and world record of 9.79 sec – for taking banned steroids, leaving Carl Lewis to take the gold. It took another 11 years for another man to equal Johnson's Olympic time – Maurice Greene, who retired in 2007 after winning world and Olympic titles, but recently had to deny accusations that he had used drugs. Meanwhile, Bolt responded to the obvious question that followed his world record by saying that he had never taken any performance-enhancing drugs.

Improvements in track surfaces and running shoes have certainly helped athletes go faster in the last 20 years, as have advances in training methods and nutrition. Nevertheless, it is generally agreed that 30mph is the likely limit for humans as things stand. What might yet push human beings beyond that limit, however, is gene therapy. As recent experiments with mice have demonstrated, this rapidly growing technology can produce profound improvements in strength, speed and endurance. It's scary stuff.

Had Jesse Owens been able to take advantage of the advances in physiology, nutrition, training, footwear and track surfaces, you fancy he would have been up there with the best in today's sprinting scene

Example:

How many times has the 100m record been broken since 1912?

18 / eighteen

1. What unofficial 'title' does Michael Johnson hold?

2. Who reached the highest speed in a race?

3. By what fraction of a second did Jessie Owens break the world record?

4. Who is the 1988 Olympics 100-metre title holder?

5. When was Ben Johnson's discredited 'record' matched?

6. What do some people suspect about Maurice Greene?

7. Name two things that have helped improve times legally.

8. What do experts believe gene therapy might affect?

9. How would Jessie Owens perform alongside today's top athletes?

(Total: 9 marks)

(Total marks for Reading: 30)

Writing Part 1
Write an email of application for the following job:

- *say why you are interested in the job*
- *say what skills you have*
- *ask whether they provide accommodation*
- *ask about pay.*

Write between 100 and 150 words.

SUNNY BAY HOLIDAY CAMP
Capable, enthusiastic and organised activity leaders needed for July and August

Are you energetic with a good sense of humour?

Do you enjoy arts and crafts?

Come and spend the summer with us

Help to organise activities for children such as:

sports tournaments, art competitions and musical performances

Apply to Mr Green (Manager): jgreen@sunnybay.com

Writing Part 2

Your friend wants you to go on holiday together but it's to a place you're not keen on. Write a ***letter*** *persuading your friend to go somewhere* ***you****'d like to go to.*

Write between 150 and 200 words.

End of Examination

Appendix 6 Example of Task Specific Scale (Task 1)

Mark Scheme B2 Writing 1					
	Task Fulfilment	Range	Organisation	Accuracy	
3 First Class Pass	Fully and appropriately addresses all four content points satisfying the demands of the task, with good expansion & support.	Broad range of Grammar & Vocabulary used with clarity, assurance and precision	Cohesive & coherent text appropriately using a full range of linguistic devices	Few if any errors of spelling or punctuation	
2 pass	Mainly satisfies the demands of the task, covering at least 3 content points with adequate expansion of the topic / content points	Range of Grammar & Vocabulary used, with no impeding errors	Cohesive & coherent text adequately using a range of linguistic devices	Some errors of spelling or punctuation, though meaning still clear	Minimum expected at B2 level
1 Narrow fail	Responds to at least 2 content points. Partially satisfying the demands of the task, with little expansion of the topic / content points	Relatively narrow range of Grammar & Vocabulary used, with some impeding errors	Attempts to use linguistic devices though not always consistent	Errors of spelling or punctuation make the text difficult to follow	
0 Clear fail	Does not satisfy the demands of the task, responding to only one or none of the content points appropriately. No expansion of the topic OR off topic	Only a rudimentary range of Grammar & Vocabulary used. Many errors, often difficult to follow	Lacks cohesion and/or uses linguistic devices inappropriately	Errors of spelling and punctuation make the text very difficult to follow	

Published by City & Guilds
1 Giltspur Street
London
EC1A 9DD
T +44 (0)20 7294 2468
F +44 (0)20 7294 2413
www.cityandguilds.com

City & Guilds is a registered charity established to promote education and training