

**Relating the Trinity College London GESE and ISE exams  
to the Common European Framework of Reference: Piloting  
of the Council of Europe draft Manual**

**Final Project Report  
February 2007**

**Spiros Papageorgiou  
Department of Linguistics and English Language  
Lancaster University**



## Table of contents

<b>ACKNOWLEDGEMENTS</b>	<b>7</b>
<b>1 INTRODUCTION</b>	<b>8</b>
1.1 Aims of the project and outline of methodology	8
1.2 Significance of the project	9
1.3 Selection of participants	10
1.4 The GESE and ISE suites	10
1.5 Structure of the present report	11
<b>2 FAMILIARISATION</b>	<b>12</b>
2.1 Methodology	12
2.1.1 Before the project meeting	12
2.1.2 During the project meeting	12
2.1.3 After the project meeting	13
2.2 Analysis of judgements	13
2.2.1 Analysis of judgements using classical statistics	14
2.2.2 Analysis of judgements using the Rasch model	16
2.2.3 Analysis of judgements from CEFR Table 3 and Manual 5.8	20
2.3 Conclusion	21
<b>3 SPECIFICATION</b>	<b>22</b>
3.1 Familiarisation tasks	22
3.1.1 Analysis of judgements using classical statistics	22
3.1.2 Analysis of judgements using the Rasch model	24
3.1.3 Conclusion	25
3.2 Methodology	26
3.2.1 Before the meeting	26
3.2.2 During the meeting	27
3.2.3 After the meeting	28
3.3 Results	28
3.4 Discussion of methodology and results	31
3.4.1 The branching approach	31
3.4.2 Fit between the CEFR descriptors and the Trinity suite	31
3.4.3 Justification and rationale behind decisions	32
3.4.4 The link between GESE and ISE	32
3.4.5 The skill of Listening	32
3.4.6 Using the Specification Forms in practice	32
3.4.7 Validity of the Specification claim	33
3.5 Conclusion	34

<b>4</b>	<b>STANDARDISATION</b>	<b>35</b>
<b>4.1</b>	<b>Familiarisation tasks</b>	<b>35</b>
4.1.1	Analysis of judgements using classical statistics	35
4.1.2	Analysis of judgements using the Rasch model	37
4.1.3	Conclusion	39
<b>4.2</b>	<b>Methodology</b>	<b>39</b>
4.2.1	Before the meeting	39
4.2.2	During the meeting	40
4.2.3	After the meeting	41
<b>4.3</b>	<b>Training</b>	<b>42</b>
4.3.1	Investigating consistency and agreement	42
4.3.2	Investigating the rating process	43
4.3.3	Conclusion	44
<b>4.4</b>	<b>Benchmarking</b>	<b>44</b>
4.4.1	Investigating consistency and agreement for the GESE suite	45
4.4.2	Investigating the rating process for GESE Initial Grades	45
4.4.3	Investigating the rating process for GESE Elementary Grades	46
4.4.4	Investigating the rating process for GESE Intermediate Grades	47
4.4.5	Investigating the rating process for GESE Advanced Grades	49
4.4.6	Investigating consistency and agreement for ISE Interview	50
4.4.7	Investigating the rating process for ISE Interview	51
4.4.8	Investigating consistency and agreement for ISE I and II Written	53
4.4.9	Investigating the rating process for ISE I and II Written	54
4.4.10	Investigating consistency and agreement for ISE 0 and III Written	57
4.4.11	Investigating the rating process for ISE 0 and III Written	57
4.4.12	Conclusion	59
<b>4.5</b>	<b>Standard-setting</b>	<b>60</b>
4.5.1	Methodology	60
4.5.2	Investigating consistency and agreement	61
4.5.3	Cut-off scores in relation to the CEFR for GESE and ISE	61
4.5.4	Discussion of the resulting cut-off scores	63
4.5.5	Considering the validity of the Standardisation claim	65
<b>4.6</b>	<b>Conclusion</b>	<b>66</b>
<b>5</b>	<b>EMPIRICAL VALIDATION</b>	<b>67</b>
<b>5.1</b>	<b>Internal Validation</b>	<b>67</b>
5.1.1	The GESE study	67
5.1.2	The ISE study	68
5.1.3	Conclusion on the GESE and ISE studies	69
5.1.4	Examiner training and its importance for the CEFR linking claim	69
5.1.5	Aims of examiner training for GESE and ISE	70
5.1.6	The Examiners' Conference	70
5.1.7	Description of the Conference programme-Day 1	70
5.1.8	Description of the Conference programme-Day 2	71
5.1.9	Description of the Conference programme-Day 3	71
5.1.10	Examiners' pack	72
5.1.11	Conclusion on examiner training	72
5.1.12	General conclusion on Internal Validation	72
<b>5.2</b>	<b>External Validation</b>	<b>73</b>
5.2.1	Indirect and direct linking: some considerations	73
5.2.2	Addressing the issue of defining a criterion	73

5.2.3	Criterion-test comparison for GESE	73
5.2.4	Criterion-test comparison for ISE	75
5.2.5	Conclusion on External Validation	76
<b>5.3</b>	<b>Conclusion on Empirical Validation</b>	<b>76</b>
<b>6</b>	<b>GENERAL CONCLUSION</b>	<b>77</b>
<b>7</b>	<b>REFERENCES</b>	<b>78</b>
	<b>Appendix 1: Familiarisation programme</b>	<b>82</b>
	<b>Appendix 2: Specification programme</b>	<b>83</b>
	<b>Appendix 3: Standardisation programme</b>	<b>84</b>
	<b>Appendix 4: Samples from the two types of Familiarisation tasks</b>	<b>85</b>
	<b>Appendix 5: Rating Form for Speaking</b>	<b>88</b>
	<b>Appendix 6: Standard Setting form for the Initial Grades</b>	<b>89</b>
	<b>Appendix 7: Ratings of samples using Trinity bands</b>	<b>90</b>

## List of Tables

<i>Table 1.1 The structure of the GESE suite</i>	10
<i>Table 1.2 The structure of the ISE suite</i>	11
<i>Table 2.1 Intra-rater reliability-summary statistics</i>	14
<i>Table 2.2 Inter-rater reliability and internal consistency-summary statistics</i>	14
<i>Table 2.3 Rater-CEFR agreement-summary statistics</i>	15
<i>Table 2.4 Scores obtained by judges in the familiarisation tasks-correct answers</i>	15
<i>Table 2.5 Rater measurement report-Familiarisation</i>	17
<i>Table 2.6 Scaling of the descriptors-round 1</i>	19
<i>Table 2.7 Scaling of the descriptors-round 2</i>	19
<i>Table 2.8 Correct placement of descriptors for CEFR Table 3 and Manual Table 5.8</i>	20
<i>Table 3.1 Intra-rater reliability-summary statistics</i>	22
<i>Table 3.2 Inter-rater reliability and internal consistency-summary statistics</i>	23
<i>Table 3.3 Rater-CEFR agreement-summary statistics</i>	23
<i>Table 3.4 Scores obtained by judges in the familiarisation tasks-correct answers</i>	23
<i>Table 3.5 Judges' characteristics- logits and infit mean square values</i>	24
<i>Table 3.6 Scaling of the descriptors</i>	25
<i>Table 3.7 Organisation of the parallel sessions for GESE and ISE</i>	27
<i>Table 3.8 Rounds of judgements during the Specification stage</i>	28
<i>Table 3.9 Relationship of GESE content to the CEFR-holistic estimation</i>	31
<i>Table 3.10 Relationship of ISE content to the CEFR-holistic estimation</i>	31
<i>Table 4.1 Intra-rater reliability-summary statistics</i>	35
<i>Table 4.2 Inter-rater reliability and internal consistency-summary statistics</i>	36
<i>Table 4.3 Rater-CEFR agreement-summary statistics</i>	36
<i>Table 4.4 Scores obtained by judges in the familiarisation tasks-correct answers</i>	36
<i>Table 4.5 Judges' characteristics- logits and infit mean square values</i>	37
<i>Table 4.6 Scaling of the descriptors</i>	38
<i>Table 4.7 Conversion of CEFR levels into quantitative data</i>	42
<i>Table 4.8 Agreement and consistency of judges-training sessions</i>	43
<i>Table 4.9 Training results-summary statistics</i>	43
<i>Table 4.10 Items from the Matriculation Examination in Finland</i>	44
<i>Table 4.11 Agreement and consistency of judges-GESE benchmarking</i>	45
<i>Table 4.12 Benchmarking Initial Grades-summary statistics</i>	46
<i>Table 4.13 Benchmarking Elementary Grades-summary statistics</i>	47
<i>Table 4.14 Benchmarking Intermediate Grades-summary statistics</i>	48
<i>Table 4.15 Benchmarking Advanced Grades-summary statistics</i>	50
<i>Table 4.16 Agreement and consistency of judges -ISE Interview benchmarking</i>	51
<i>Table 4.17 Benchmarking ISE 0 Interview-summary statistics</i>	51
<i>Table 4.18 Benchmarking ISE I Interview-summary statistics</i>	52
<i>Table 4.19 Benchmarking ISE II Interview-summary statistics</i>	53
<i>Table 4.20 Benchmarking ISE III Interview-summary statistics</i>	53
<i>Table 4.21 Agreement and consistency of judges-ISE I and II Written benchmarking</i>	54
<i>Table 4.22 Benchmarking ISE I Written-summary statistics</i>	55
<i>Table 4.23 Benchmarking ISE II Written-summary statistics</i>	56
<i>Table 4.24 Agreement and consistency of judges-ISE 0 and III Written benchmarking sessions</i>	57
<i>Table 4.25 Benchmarking ISE 0 Written-summary statistics</i>	58
<i>Table 4.26 Benchmarking ISE III Written-summary statistics</i>	59
<i>Table 4.27 Agreement and consistency of GESE cut-scores judgements</i>	61
<i>Table 4.28 Cut-off scores in relation to the CEFR-GESE round 1</i>	62
<i>Table 4.29 Cut-off scores in relation to the CEFR-GESE round 2</i>	63
<i>Table 4.30 Cut-off scores in relation to the CEFR-ISE</i>	63
<i>Table 4.31 CEFR level of borderline and secure pass candidates in the GESE suite</i>	64
<i>Table 4.32 CEFR level of borderline candidates in the ISE suite</i>	65
<i>Table 5.1 Examiners-monitors scoring agreement for ISE</i>	68
<i>Table 5.2 CEFR level comparison for GESE</i>	74
<i>Table 5.3 CEFR level comparison for ISE</i>	75

## List of Figures

<i>Figure 2.1 Ruler map for the speaking 1 task</i>	17
<i>Figure 3.1 Graphic Profile of the GESE-CEFR relationship (Initial and Activities)</i>	29
<i>Figure 3.2 Graphic Profile of the GESE-CEFR relationship (Competences)</i>	29
<i>Figure 3.3 Graphic Profile of the ISE-CEFR relationship (Initial and Activities)</i>	30
<i>Figure 3.4 Graphic Profile of the ISE-CEFR relationship (Competences)</i>	30
<i>Figure 5.1 CEFR Decision Table for GESE</i>	74
<i>Figure 5.2 CEFR Decision Table for ISE</i>	75

## **Acknowledgements**

This project has benefited from the work and feedback of a number of people to whom I am grateful.

Charles Alderson, Neus Figueras and Felianka Kaftandjieva have shared with me their experience in using the Framework in language testing and have always been happy to reply to my queries about the linking process of the Manual.

The 12 project participants worked very hard in order to build a good understanding of the CEFR that would result in meaningful and reliable judgements about the relevance of the Trinity exams to the CEFR. I feel the need to thank them not only for their hard work but also for tolerating me as a very demanding and strict project coordinator.

Apart from the 12 project participants, a number of people from Trinity College London contributed by providing administrative support for this project, which was essential for building a valid and reliable claim about the CEFR linkage. I am grateful also to them for making me feel at home whenever I visited the Head Office for the project meetings and at the same time for ensuring that the project meetings were carried out in the most professional atmosphere.

Finally, many thanks go to all those who have contributed with thought-provoking comments in a number of conferences where the project has been presented and especially the members of the Language Testing Research Group at Lancaster University. Naturally, any flaws in the present report rest with the author.

# 1 Introduction

The *Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (Council of Europe, 2001), known as CEF or CEFR, has become the most influential project of the Council of Europe on teaching, curriculum design, learning and assessment. In the first detailed collection of articles employing the CEFR in a number of areas in education (Alderson, 2002b) the editor notes:

Clearly the influence of the Framework has been widespread and deep, impacting on curricula, syllabuses, teaching materials, tests and assessment systems and the development of scales of language proficiency geared to the six main levels of the CEFR.  
(Alderson, 2002a: 8)

The CEFR provides a common basis for the description of objectives, content and methods intending to “enhance the transparency of courses, syllabuses and qualifications, thus promoting international co-operation in the field of modern languages” (Council of Europe, 2001:1).

With specific reference to language testing, following the publication of the CEFR it became apparent that language tests had a common reference point, that is, the set of six levels. Therefore, language tests could be compared easily by referring to learners who would sit the exam as B1, B2 etc. Transparency among language qualifications and comparability seemed to be plausible. But how would reference to the CEFR levels be achieved following good practice and empirical evidence, without superficially intuitive understanding of the level of the learners? Applying only intuitive criteria as to the way tests relate to the CEFR levels would obviously result in invalid claims. How then would test constructors validly claim that their tests examine language used by B2 learners? A methodology for relating tests to the CEFR was needed in order for transparency among language qualifications to be achieved and valid claims as to the relation of tests to the CEFR to be made.

The response of the Council of Europe to that need was the publication in 2003 of a pilot version of the *Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (Council of Europe, 2003) along with a Reference Supplement, all available from the Council’s website ([www.coe.int/portfolio](http://www.coe.int/portfolio)). A description of the Manual and its process of linking exams to the CEFR can also be found in Figueras et al. (2005) and North (2004).

After the publication of the Manual, the Council of Europe invited exam providers to pilot it and provide feedback on the linking process, aiming at the revision of the Manual, the production of calibrated samples to the CEFR and the publication of a case studies book. In total, 40 institutions from 20 countries are currently participating in the piloting of the Manual (Martyniuk, 2006). Trinity College London is among the exam boards participating in the piloting of the Manual. The present report describes the methodology and the results of this project commissioned in February 2005.

## 1.1 Aims of the project and outline of methodology

The Trinity CEFR calibration project aims at mapping and standardising GESE and ISE suites to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (Council of Europe, 2001). The Manual for relating exams to the CEFR (Council of Europe, 2003) describes the methodology for the

linking process which was followed in the project. The linking process comprises four sets of interrelated activities:

- 1. Familiarisation.** This stage, which should be repeated before Specification and Standardisation, is imperative in order to ensure that the members of the linking panel are familiar with the content of the CEFR and its scales. Familiarisation tasks are suggested by the Manual. In relation to the Trinity project, this stage took place as a separate phase on 6-7 September 2005 and will be discussed in section 2. The venue for all phases was the Trinity Head Office in London.
- 2. Specification.** This stage involves the description of the content of the test to be related to the CEFR first in its own right and then in relation to the levels and categories of the CEFR. Forms for the mapping of the test are provided by the Manual. The outcome of this stage is a claim regarding the content of the test in relation to the CEFR. The Trinity Specification stage took place on 22-24 November 2005 and is discussed in section 3.
- 3. Standardisation.** The outcome of this stage is the reinforcement of the previous claim. Standardisation involves achieving a common understanding of the CEFR levels illustrated by examples of actual learners' performance. Standardisation techniques are offered by the Manual. The Trinity Standardisation took place from 28 February to 2 March 2006 and is discussed in section 4.
- 4. Empirical validation.** There are two categories of empirical validation in the Manual. Internal validation aims at establishing the quality of the test in its own right. External validation aims at the independent corroboration of the standards set by either using an anchor test already calibrated to the CEFR, or by using judgements of teachers well trained in the CEFR. The outcome of this stage is the confirmation or not of the claims in the two previous stages by using analysed test data. Internal and external validations are described in section 5 of the present report.

## 1.2 Significance of the project

The results of this project are of interest to a number of parties. First, the Council of Europe can obtain feedback on the piloting of the Manual and build on it for the next version of the Manual which will follow the preliminary draft one, currently available.

The project is of course beneficial for Trinity; it is not only the wide acceptance that Trinity qualifications could have after being related to the CEFR, but also, given that this project is combined with extensive Empirical Validation and external and internal monitoring of the Trinity exams, confidence in the quality of these exams will be even higher. The project may also be used by Trinity as an impetus to further research and on-going improvements to the examinations and examiner standardisation.

Test users, such as candidates, parents, teachers, employers and educational institutions will also benefit from the results of the project. This is because the project aims at explaining to test users what a score means when taking a Trinity exam in the CEFR terminology. Because the CEFR has become the common language in Europe, such a description of test scores is essential.

The above points are directly relevant to the comparability and transparency aims of the Council of Europe and one of the primary intentions of the Manual which is awareness-raising of good testing practice and quality of the tests using the CoE

documentation. Finally, this project can potentially contribute to research in the area of language testing, as it the main source of data for the author’s doctoral study, currently in progress at Lancaster University.

### 1.3 Selection of participants

The quality of judgements is a vital point as is frequently pointed out in the relevant literature (Kaftandjieva, 2004:4), linking to the CEFR, like standard setting which is one of its components, is a highly judgemental and arbitrary process, but does not need to be capricious (Glass, 1978; Popham, 1978). In order to avoid capriciousness, well-designed methodology and training of judges are imperative.

In order to select judges for the project all the above were considered. Therefore in July 2005 I liaised with the Trinity Chief Examiner in order to choose and invite 12 judges (the Manual suggests at least 10) who would have a variety of responsibilities and roles. The judges chosen were involved in examining, marking, monitoring, validation, test design and administration of Trinity tests. Two of them were managers in the ESOL department. The group overall was familiar to some extent with the CEFR, as it had already been used in various stages of the design of the Trinity tests such as marking criteria and content specifications.

The names of the 12 panellists are not revealed here for confidentiality and ethical reasons related to the use of the recordings from the meetings for the purposes of the author’s doctoral research. Pseudonyms will be used throughout the report. More details on the characteristics of the panel can be obtained from the author.

### 1.4 The GESE and ISE suites

According to the syllabus of each exam (Trinity College London, 2005a, 2005b), the Graded Examinations in Spoken English examination suite tests speaking and listening during a one-to-one interaction with an examiner and has 12 levels from Grade 1 to Grade 12 (see Table 1.1).

*Table 1.1 The structure of the GESE suite*

<b>Initial</b>	<b>Elementary</b>	<b>Intermediate</b>	<b>Advanced</b>
Grades 1-3 5-7 mins	Grades 4-6 10 mins	Grades 7-9 15 mins	Grades 10-12 25 mins
			Topic presentation Topic discussion
		Topic presentation and discussion	Listening Task
	Topic Discussion Conversation	Interactive Task Conversation	Interactive Task Conversation
Conversation			

The Integrated Skills in English suite of examinations follows the same structure for the spoken component but there are two additional components, a portfolio and a controlled written exam which test writing and reading in an integrated way (see Table 1.2).

*Table 1.2 The structure of the ISE suite*

<b>ISE levels</b>	<b>Components</b>
ISE 0	3 portfolio tasks controlled written examination an oral interview
ISE I	
ISE II	
ISE III	

## **1.5 Structure of the present report**

The structure of the report follows the order of the Chapters in the Manual and the chronological order according to which each phase of the project took place. Section 2 reports on the Familiarisation phase, section 3 discusses Specification, and finally, the Standardisation phase is the theme of discussion in section 4. Empirical Validation is described in section 5.

## **2 Familiarisation**

In this section I present statistical analysis of the two-day-long Familiarisation meeting in September 2005. Following the Manual, Familiarisation tasks aimed at securing in-depth understanding of the CEFR scaled descriptors, because these are the main instruments to be used for the consequent linking stages. The research question explored was set as following:

*Are the judges consistent when using the CEFR scales and do they have a deep understanding of the CEFR levels?*

The methodology of the project and the results of the Familiarisation tasks are discussed in detail in the following sections.

### **2.1 Methodology**

The methodology of the Familiarisation is explained here in three sections, following chronological order. The programme of the meeting is included in Appendix 1 (p. 82).

#### **2.1.1 Before the project meeting**

The panellists were asked to study the CEFR volume and familiarise themselves with the scaled descriptors prior to the Familiarisation meeting. A booklet containing CEFR descriptors was prepared. This was first piloted in Lancaster, UK with informants who did the same tasks as the Trinity judges.

The booklet contained the following Common Reference Levels descriptors from Table 2 of the CEFR (Council of Europe, 2001: 26-27): 30 speaking, 25 writing, 19 listening and 20 reading descriptors. From Table 1 (Council of Europe, 2001: 24) 30 global descriptors were included. The same number of descriptors was chosen for qualitative aspects of spoken language use from Table 3 (Council of Europe, 2001: 28-29), as well as 28 descriptors from the Manual's written assessment criteria grid in Table 5.8 (Council of Europe, 2003: 82).

The writing, listening and reading descriptors from Table 2 were taken from Kaftandjieva and Takala (2002), who have broken down the original descriptors into their constituent sentences. Following this approach, speaking and global descriptors were created. The use of smaller descriptors was very interesting from a research point of view, because it could provide insights as to whether sentences belonging to the same descriptors were assigned the same level. However, as judges argued at the end of the Familiarisation meeting, this approach made level guessing much more complicated, which should be taken into account when results are discussed later.

#### **2.1.2 During the project meeting**

Judges were asked to guess and write next to the statements from the first five scales (from CEFR Tables 1 and 2) the CEFR level of the descriptor (A1-C2). For the last two sets of descriptors from CEFR Table 3 and Manual Table 5.8 the judges were given the descriptors in small confetti-style pieces and were asked to stick them on the cells they belonged to. The effect that the different task format could have on the accuracy of judgements was considered during the piloting of the instruments in Lancaster and there was no clear evidence that the different format affected judgements. Appendix 4 (p. 85) contains samples of the two tasks types.

The coordinator used a laptop and a data projector in the room and after the accomplishment of the task for a descriptor set, each judge indicated the level he/she chose and that level was displayed later on the screen using EXCEL. After all judges reported the level of a descriptor, the correct level of the descriptor appeared on the screen and discussion followed regarding the reasons for choosing a particular level. This process was repeated for all descriptors. The discussions were all recorded for further analysis and clarification of comments by the panel. Placement of the first 5 sets of descriptors from CEFR Tables 1 and 2 were repeated on the second day in order to investigate intra-rater reliability.

For the confetti-style descriptors, judges were asked to fill in the cells individually and then discuss in pairs their answers. The correct answers were provided after the paired discussion and a group discussion followed, in which the pairs reported on their experience of this task

### **2.1.3 After the project meeting**

All tasks resulted in 3672 level placements and each candidate received by email the number of correct placements he/she achieved for all tasks. In order to investigate panellists' consistency, levels were converted into numbers (A1=1, A2=2, etc) and judgements were analysed for intra- and inter-rater reliability and internal consistency, following Kaftandjieva and Takala (2002). Spearman rank correlation coefficient is reported for rater reliability because of the assumptions made for the data, which only need to constitute an ordinal scale (Bachman, 2004:88). The Pearson product-moment correlation coefficient which is very frequently encountered in the literature and is also reported in the Kaftandjieva and Takala (2002) study, requires further assumptions to be made, and because it produces more or less similar results to Spearman as was realised in the GESE/ISE analysis of judgements, it is not reported here since it will not add any more information. The average of these correlation coefficients is calculated using Fisher's Z-transformation (Bachman, 2004:170).

Cronbach  $\alpha$  (alpha) is an internal consistency index usually used in item-based tests, but following studies analysing similar familiarisation tasks (Generalitat de Catalunya, 2006; Kaftandjieva & Takala, 2002), I will also report it here as an indication of "the consistency of the reliability of ratings in terms of rater consistency" (Generalitat de Catalunya, 2006:62).

Finally, many-facet Rasch measurement was employed using the FACETS programme version 3.58 (Linacre, 2005). The aim of this analysis was to investigate whether the group could scale the descriptors according to the expected pattern from lower level to higher level descriptors.

## **2.2 Analysis of judgements**

Analysis of judgements is divided into three parts. First in subsection 2.2.1 I report on analysis using classical statistics, whereas analysis using the Rasch model is discussed in subsection 2.2.2. This distinction in the analysis of judgements is based on the classification of the Rasch model under the one parameter Item Response Theory model. Such models are based on a measurement theory called Item Response Theory (Hambleton, Swaminathan, & Rogers, 1991), which offers some advantages over classical item analysis. The analysis of judgements will therefore be carried out using both theories of measurement. Finally, in subsection 2.2.3 judgements for confetti-style tasks from CEFR Table 3 and Manual 5.8 are discussed.

### 2.2.1 Analysis of judgements using classical statistics

In this section I report on the findings regarding intra-rater reliability, inter-rater reliability and agreement with the CEFR scales. Cohen et al. (2000) consider coefficients above .65 satisfactory in the field of education, whereas in the Kaftandjieva and Takala (2002) study coefficients above .7 are reported as satisfactory. In large scale assessments, inter-rater reliability is usually expected to be even higher, in the area of .8 (Alderson, Clapham, & Wall, 1995:132). For reasons explained later, I will concentrate on the analysis of judgements for descriptors in Tables 1 and 2 of the CEFR.

Table 2.1 presents a summary of intra-rater correlations for the set of descriptors from CEFR Tables 1 and 2. The correlation was calculated by comparing the first and the second time each judge assigned a CEFR level to a set of descriptors, thus showing how consistent the panellists were with themselves. Spearman correlations in Table 2.1, as well as all other tables in the present report are statistically significant ( $p \leq 0.01$ ), which means that we can be 99% sure that correlations did not occur by chance. Intra-rater reliability is very high for all five descriptor sets.

*Table 2.1 Intra-rater reliability-summary statistics*

Scales	Intra-rater reliability		
	Mean*	Min	Max
Speaking	0.915	0.896	0.932
Writing	0.879	0.748	0.926
Listening	0.919	0.845	0.971
Reading	0.951	0.894	0.994
Global	0.948	0.872	0.969

\*Average using Fisher's Z-transformation

Table 2.2 presents high levels of inter-rater reliability, that is, agreement between two judges. Inter-rater reliability was calculated by running Spearman correlations for all possible pairs of judges in order to investigate whether they agreed on ranking the descriptors from the lowest to the highest level. With the exception of some lower minimum values for Writing and Listening, inter-rater reliability is high and this is supplemented by the high Alpha index for all sets of descriptors.

*Table 2.2 Inter-rater reliability and internal consistency-summary statistics*

Scales	Inter-rater reliability			Alpha
	Mean*	Min	Max	
Speaking1	0.894	0.831	0.958	0.958
Speaking 2	0.934	0.877	0.966	0.966
Writing 1	0.866	0.751	0.939	0.939
Writing 2	0.868	0.719	0.955	0.955
Listening 1	0.915	0.820	0.975	0.991
Listening 2	0.891	0.718	0.963	0.990
Reading 1	0.932	0.837	0.979	0.991
Reading 2	0.942	0.887	0.988	0.994
Global 1	0.921	0.844	0.963	0.992
Global 2	0.937	0.846	0.977	0.994

\*Average using Fisher's Z-transformation

Table 2.3 presents the agreement between the panellists' level assignment with the correct level. Spearman correlations were run between each judge's levels and the correct CEFR levels. These correlations are all very high. It should be stressed here that Spearman coefficient shows rank order correlations, which in this context means that it explains agreement in the order that two sets of descriptors had been arranged and should not be interpreted as exact agreement of assigned levels. Even a correlation of 1 can occur with 0% exact agreement if different ranges of the scale are used as pointed out by Kaftandjieva (2004:24); for this reason, another coefficient is included: Cohen's  $\kappa$  (Kappa) calculates exact agreement by also taking into account agreement by chance, which cannot be taken into account when reporting raw scores. Kappa is reported in Table 2.3 along with Spearman correlations.

Table 2.3 Rater-CEFR agreement-summary statistics

Scales	Spearman correlations			Cohen's Kappa			N
	Mean*	Min	Max	Mean	Min	Max	
Speaking 1	0.911	0.871	0.928	0.464	0.28	0.602	10
Speaking 2	0.958	0.913	0.985	0.626	0.282	0.88	12
Writing 1	0.883	0.791	0.938	0.423	0.228	0.516	11
Writing 2	0.907	0.828	0.957	0.547	0.335	0.709	10
Listening 1	0.907	0.832	0.961	0.548	0.408	0.74	11
Listening 2	0.920	0.855	0.962	0.593	0.422	0.805	12
Reading 1	0.959	0.901	1	0.591	0.235	1	11
Reading 2	0.968	0.923	0.994	0.687	0.474	0.939	12
Global 1	0.939	0.901	1.000	0.589	0.36	0.84	12
Global 2	0.959	0.923	0.994	0.66	0.439	0.88	12

\*Average using Fisher's Z-transformation

The  $\kappa$  coefficient shows that exact agreement is not as high as the rank order correlation. This probably means that the judges can understand in general lower and higher levels, but due to the large number of descriptor units, some of them mixed adjacent levels. For this reason, the discussion that followed the task, during which judges could see their rating on the projector, aimed at helping them to see the differences between such levels. A disadvantage of Kappa is that it presupposes that the values of the two variables match, otherwise the coefficient cannot be calculated. In the present study this means that Kappa cannot be calculated if a judge has not used all six CEFR levels; this is why the last column in Table 2.3 shows the number of judges whose ratings could be calculated because they used the whole range of levels.

Even though raw scores do not take into account agreement by chance as is the case with coefficient  $\kappa$ , these are included in Table 2.4, because they provide some further insights into the actual rating process, that is, they show how many descriptors were placed at the correct level. The data was treated here in a dichotomous way, that is, 0 for wrong level placement and 1 for correct one. As already stated previously, exact agreement is not very high, but it is encouraging that results are better in the second round for each set, suggesting that group discussion had a positive effect on the judges' understanding of the scales.

Table 2.4 Scores obtained by judges in the familiarisation tasks-correct answers

Scales	Descriptors	Mean	Min	Max	SD
Speaking 1	30	16.5	12	20	2.78
Speaking 2	30	20.67	12	27	4.29
Writing 1	25	12.5	6	15	2.75
Writing 2	25	14.83	10	19	3.16
Listening 1	19	11.58	8	15	1.93
Listening 2	19	12.5	10	16	1.78
Reading 1	20	12.92	7	20	3.99
Reading 2	20	14.67	10	19	2.9
Global 1	30	19.83	14	26	3.33
Global 2	30	21.5	16	27	3.42

Overall, classical statistics show high levels of consistency, with some reservations as to whether judges can distinguish between adjacent levels as  $\kappa$  coefficient and raw scores show. Given however the number of descriptors and the process of mutilating the original statements into smaller units, this is hardly surprising.

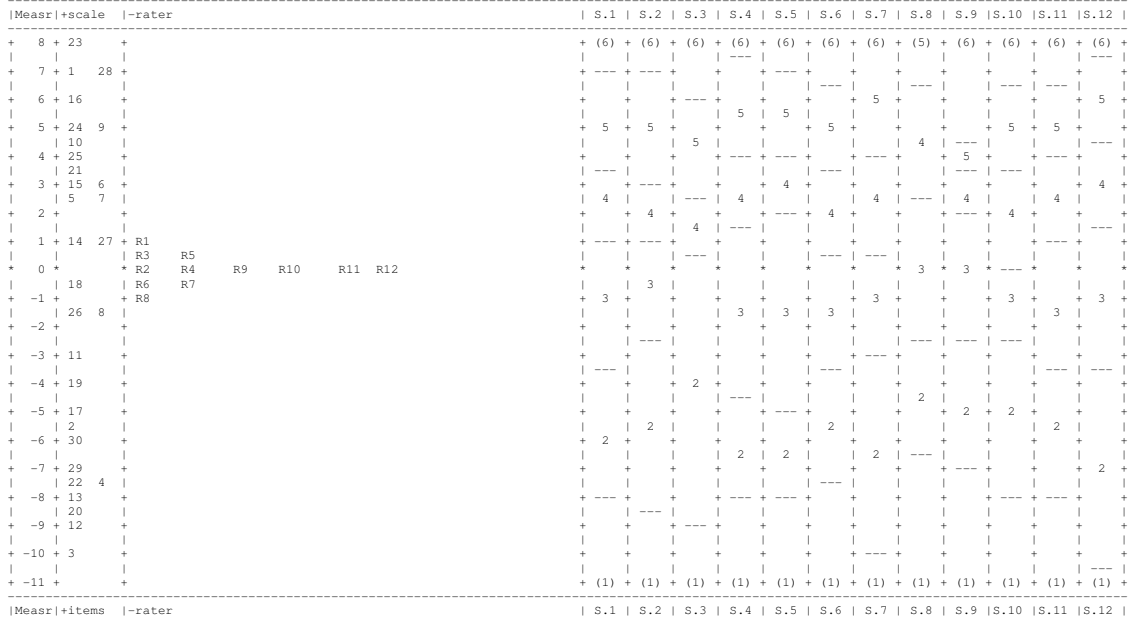
### 2.2.2 Analysis of judgements using the Rasch model

The FACETS programme used in this study (Linacre, 2005) enables the use of the many-facet Rasch model (Linacre, 1989) an extension of the earlier Rasch models such as the basic Rasch model for dichotomous data, the Rating Scale model (Andrich, 1978) and the Partial Credit model (Masters, 1982). The many-facet Rasch model can include a number of facets involved in assigning a score. The two facets in the earlier models are items and persons, whereas the many-facet model allows for other factors to be included in the measurement process. For example in performance assessment, raters, examiners and tasks could be included in the analysis. The advantage of such analysis is that all these facets can be compared on a common metric scale. In the present study this is illustrated in Figure 2.1, in which the ruler map for the first round of the speaking can be seen.

The first column, measurement, is an arbitrary interval scale separated in logits units. The scale is centred around 0. This makes more sense if we consider the two facets of the speaking task. The column entitled “scale” shows the items, which in this context are the CEFR descriptors. The second column shows the raters, that is the panellists. We can see that the descriptors are spread across a wide range of logits. Descriptors at the higher end of the scale are considered more difficult than those at the lower end; therefore, Trinity judges estimated that item 3 (that is the descriptor labelled S3 in the judges’ handout) is the easiest item (see second column “scale”) and item 23 (that is descriptor labelled S23 in the judges’ handout) the most difficult.

The raters are also measured on the same scale. In the rater column we see how the judges differ in terms of harshness and leniency. Raters are negatively orientated as opposed to the scale facet which is positively orientated as denoted by ‘-’ and ‘+’ next to the name of each facet (Linacre, 2005:93). A rater higher on the scale is stricter than the others, which in the present study is interpreted as assigning lower levels to the descriptors compared to other raters. The rater column shows that there are not notable differences in estimating the difficulty of the descriptors: all judges are within two logits (-1 to +1). The S1-S12 columns also show how each individual used the scale and confirm that as already stated above, some raters do not use all levels. For example, we see that Judge 2 has a notable difference in relation to the others: there is no level 3, i.e. B1.

Figure 2.1 Ruler map for the speaking 1 task



Ruler maps for all descriptors sets generated similar results, that is descriptors are spread out across a large number of logits and judges do not appear to differ in terms of leniency/severity. Further aspects of judges' characteristics are summarised in Table 2.5.

Table 2.5 Rater measurement report-Familiarisation

Scales	logits (min-max)	Infit mean square range	Reliability
Speaking 1	-.97-.83	.54-1.11	.43
Speaking 2	-1.40-2.07	.31-1.31	.71
Writing 1	-1.48-1.63	.50-1.24	.73
Writing 2	-1.20-1.47	.47-1.28	.66
Listening 1	-1.43-1.15	.19-1.27	.52
Listening 2	-1.10-1.82	.31-1.72	.49
Reading 1	-2.00-1.44	.28-1.34	.76
Reading 2	-3.00-3.17	.33-1.44	.88
Global 1	-1.03-.72	.46-1.55	.52
Global 2	-.61-.87	.47-1.81	.36

In column 1 of Table 2.5 the actual descriptor set is defined along with the corresponding round. Column 2 presents the minimum and maximum logit values of the raters on the logit scales, visually represented in Figure 2.1 under the rater column. Trinity judges are in general aligned as far as severity is concerned. The highest spread of raters is observed in Reading 2 and the lowest in Global 2.

Fit statistics are fundamental in Rasch analysis and can be observed in column 3. The Rasch model starts with an observation of an initial set of data and generates a prediction for each observation. The estimation procedure is called calibration and

each successive, recursive cycle of estimation is called iteration (McNamara, 1996:162). After multiple iterations, the model reaches a required level of accuracy between the expected and the observed estimation for all data. If the observed values do not fit the expectations of the model, then these values are misfitting. A number of fit statistics are provided by FACETS, the most meaningful of which is infit mean square (column 3). A value of 1 means that the observed value fits the expected one; values above 1 signal greater variation than expected and below 1 less variation than expected. A rule of thumb according to McNamara is that for test items values above 1.3 are significant misfit, thus there is lack of predictability, a value below .75 is significant overfit, meaning lack of variation. This use of data fit differentiates Rasch from other IRT models where parameters such as item discrimination and guessing are manipulated to maximise the fit of the model to the data (Bond & Fox, 2001:191).

In terms of interpreting rater fit values in Table 3, high fit values (above 1.3) might mean that there is lack of consistency. If we have low fit values (below .75), that might mean that the raters are consistent, but perhaps not using the full range of the scale (McNamara, 1996:139). Therefore, the low fit values in Writing, Reading and Listening could be due to the fact that the raters are very consistent or because the descriptor sets did not include many A1 descriptors in Writing and many C2 descriptors in Listening and Reading, then some raters appear not to use the whole range of levels. Consequently, the low fit values should not be a problem because it probably means that raters produce less variance than expected. Avoiding using some levels does not appear to be the reason for overfit of raters; one of the very few cases where this happens can be seen in Figure 2.1, where R3 is not using B1, that is, level 3 is not present in the S3 column.

There are, however, some high fit values at least if we consider 1.3 as the borderline. In order to interpret the results, we should consider how FACETS was used here. In this context, judges predict the level of the items (i.e. descriptors) but we should bear in mind that there is a correct answer for the items, that is, the level of each CEFR descriptor. However, the data has been inserted in the programme in a simple, two-facet design, without any indication of the correct levels. Therefore, a high infit value shows that a judge has used the levels in a different way compared to the other judges, but it does not reveal who chose the correct level. In the relevant literature, many authors note that there is no single explanation when it comes to interpretation of fit values; for example Weigle (1998:276) considers .5-1.5 an acceptable range.

In Table 2.5 we see that only the Global scale includes infit values above 1.5. In Global 1 the only high value observed, 1.55, is because Claudia in two cases assigned a score that was different from other raters; in two cases she chose A2 and B2 when everyone else chose B1 and C1/C2 respectively. Lora was the only misfitting person in Global 2, with 1.81 logits, most probably because she assigned C2 when all the other chose B2 and C1.

The final column of Table 2.5 shows another important statistic. It should be stressed that this reliability index does not show consistency of the raters but the reliable difference in severity of the raters. In other words, if this index is high then there is real difference between raters in the actual levels of scores assigned by them to candidates (McNamara, 1996:140); therefore a low reliability index is desired for the rater facet. In McNamara's example a reliability index of .99 is interpreted as indicative of real differences in raters' severity. Weigle (1998:277) also found that even after training reliability for the raters in her study was still high, at .91. The Trinity judges present moderate to low indices compared to these figures, with the

exception of Reading 2. In Reading 2 the high index is illustrated by the wide range of the raters logit as can be seen in column 2 of Table 2.5.

Overall, analysis of judgements suggests consistency of the panellists, an argument supported by the acceptable range of infit values and relatively low reliability index. A final interesting finding is illustrated in Table 2.6.

Table 2.6 Scaling of the descriptors-round 1

Speaking			Writing			Listening			Reading 1			Global		
Descr.	Logit value	CEFR level	Descr.	Logit value	CEFR level	Descr.	Logit value	CEFR level	Descr.	Logit value	CEFR level	Descr.	Logit value	CEFR level
S3	-10.13	A1	W25	-12.11	A1	L4	-12.52	A1	R6	-15.12	A1	G12	-10.10	A1
S12	-8.94	A1	W24	-10.07	A1	L3	-10.23	A1	R19	-13.78	A1	G13	-10.10	A1
S20	-8.49	A1	W6	-8.47	A2	L6	-9.57	A1	R4	-9.83	A1	G20	-8.88	A1
S13	-8.08	A1	W19	-7.81	A2	L12	-6.99	A2	R12	-9.83	A2	G3	-8.19	A1
S4	-7.30	A2	W4	-5.51	A2	L16	-6.60	A2	R16	-9.83	A2	<b>G29</b>	<b>-8.19</b>	<b>A2</b>
S22	-7.30	A2	W16	-3.86	B1	L15	-4.90	A2	R15	-6.94	A2	<b>G30</b>	<b>-7.85</b>	<b>A1</b>
S29	-6.90	A2	W12	-2.52	B1	L11	-1.79	B1	R11	-2.71	B1	G17	-7.10	A2
<b>S30</b>	<b>-6.02</b>	<b>A1</b>	W15	-2.52	B1	L5	-1.45	B1	R7	-1.63	B1	G4	-6.68	A2
<b>S2</b>	<b>-5.54</b>	<b>B1</b>	W11	-1.84	B1	L7	-.15	B2	R8	-.67	B1	G22	-6.68	A2
<b>S17</b>	<b>-5.06</b>	<b>A2</b>	W8	-.56	B2	L8	2.06	B2	R17	-.67	B1	G11	-6.24	A2
<b>S19</b>	<b>-4.20</b>	<b>B1</b>	W7	2.36	B2	L2	3.71	B2	R5	.62	B1	G19	-3.42	B1
<b>S11</b>	<b>-3.06</b>	<b>A2</b>	W22	2.74	B2	L13	4.04	B2	R2	5.06	B2	G26	-2.97	B1
S8	-1.50	B1	W5	3.43	B2	L18	4.04	C1	R13	5.94	B2	G2	-1.69	B1
S26	-1.50	B1	W21	3.98	C1	L10	4.37	C1	<b>R18</b>	<b>5.94</b>	<b>C1</b>	G18	-.51	B1
S18	-.26	B1	W17	4.22	C1	<b>L17</b>	<b>5.42</b>	<b>B2</b>	<b>R20</b>	<b>6.64</b>	<b>B2</b>	G27	-1.14	B1
<b>S14</b>	<b>1.06</b>	<b>B2</b>	W14	4.83	C2	<b>L19</b>	<b>6.17</b>	<b>B2</b>	R9	7.48	C1	G8	.55	B1
S27	1.06	B1	W9	5.19	C2	L9	7.24	C1	R1	7.71	C1	<b>G9</b>	<b>2.60</b>	<b>C1</b>
S5	2.53	C1	W1	5.54	C2	L14	9.14	C1	R10	9.03	C1	G14	2.88	B2
<b>S7</b>	<b>2.74</b>	<b>B2</b>	W3	5.54	C2	L1	10.62	C2	R14	9.03	C1	G28	3.15	B2
S15	2.95	C1	<b>W2</b>	<b>5.72</b>	<b>C1</b>				R3	11.86	C2	G6	4.48	B2
S6	3.15	B2	<b>W10</b>	<b>5.90</b>	<b>C2</b>							G6	4.48	B2
S21	3.36	B2	W13	6.51	C1							G21	4.48	B2
S25	4.20	C1	W20	6.51	C1							G15	5.32	C1
<b>S10</b>	<b>4.68</b>	<b>C2</b>	W18	6.77	C2							G5	6.21	C1
S9	4.94	C1	W23	7.07	C2							<b>G23</b>	<b>6.21</b>	<b>C2</b>
S24	5.22	C1										<b>G24</b>	<b>6.82</b>	<b>C1</b>
S16	5.82	C2										<b>G10</b>	<b>8.49</b>	<b>C2</b>
S28	6.80	C2										<b>G25</b>	<b>8.93</b>	<b>C1</b>
S1	7.16	C2										G1	9.49	C2
S23	10.63	C2										G16	10.34	C2

In Table 2.6 one can see a summary of how the group has ordered the descriptors. This ordering is based on the logit values of descriptors. In general, the pattern appears to be very close to the correct one. Highlighted levels in the third column indicate problematic scaling for example S2 (B1 level) has a smaller value than S17 (A2 level). In other words the correct scaling of the descriptors should be according to their rank order in the third column of each set: correct scaling starts with A1 descriptors, then A2 should follow, then B1 and so on. Even though scaling is generally very similar to the expected pattern, some wrong ordering of the descriptors reveals the importance of the group discussion after the task, where judges received feedback on their decision-making and wrong level placements were pointed out. Table 2.7 shows that the effect of group discussion was positive as distorted scaling is minimal in the second round.

Table 2.7 Scaling of the descriptors-round 2

Speaking			Writing			Listening			Reading 1			Global		
Descr.	Logit	CEFR	Descr.	Logit	CEFR	Descr.	Logit	CEFR	Descr.	Logit	CEFR	Descr.	Logit	CEFR

	value	level		value	level		value	level		value	level		value	level
S20	-12.13	A1	W25	-13.50	A1	L4	-11.79	A1	R6	-14.61	A1	G12	-11.97	A1
S13	-11.49	A1	W24	-12.13	A1	L3	-11.78	A1	R19	-14.61	A1	G13	-11.97	A1
S30	-10.98	A1	W19	-9.28	A2	L6	-10.46	A1	R4	-12.38	A1	G3	-10.15	A1
S3	-10.10	A1	W6	-8.36	A2	L15	-6.55	A2	R16	-10.70	A2	G30	-10.15	A1
S4	-9.26	A2	W4	-6.61	A2	L16	-6.55	A2	R12	-10.19	A2	G20	-9.75	A1
<b>S12</b>	<b>-8.34</b>	<b>A1</b>	W16	-4.52	B1	L12	-5.77	A2	R15	-8.43	A2	G11	-8.66	A2
S22	-7.84	A2	W12	-3.14	B1	L11	-1.35	B1	R11	-4.61	B1	G4	-8.29	A2
S29	-6.82	A2	W15	-2.07	B1	L5	1.06	B1	R7	-4.10	B1	G29	-8.29	A2
S17	-5.45	A2	W11	-0.66	B1	L7	1.06	B2	R8	-3.59	B1	G17	-7.51	A2
S19	-4.10	B1	W8	.08	B2	L8	2.55	B2	R17	-2.52	B1	G22	-5.67	A2
S11	-3.63	A2	W5	3.70	B2	L2	3.66	B2	R5	-1.98	B1	G26	-3.87	B1
S2	-3.14	B1	W7	3.70	B2	L13	3.92	B2	R2	3.57	B2	G19	-3.42	B1
S8	-1.61	B1	W22	3.70	B2	L17	3.92	B2	R13	3.58	B2	G2	-1.95	B1
S27	-1.61	B1	W2	5.72	C1	L19	4.18	B2	R20	5.94	B2	G18	-1.44	B1
S26	-1.11	B1	W17	6.01	C1	L18	4.43	C1	R18	6.65	C1	G27	-1.44	B1
S18	-0.20	B1	W13	6.52	C1	L10	5.13	C1	R1	7.97	C1	G8	-0.53	B1
S14	2.56	B2	W20	6.52	C1	L9	5.36	C1	R9	7.97	C1	G14	2.66	B2
S21	2.88	B2	W3	6.76	C2	L14	6.88	C1	R10	8.63	C1	G28	3.35	B2
S6	3.20	B2	W14	7.00	C2	L1	9.51	C2	R14	10.52	C1	<b>G9</b>	<b>4.04</b>	<b>C1</b>
S7	3.20	B2	W18	7.46	C2				R3	12.70	C2	G6	5.00	B2
S15	4.16	C1	W10	7.70	C2							G7	5.00	B2
S5	4.78	C1	<b>W21</b>	<b>7.70</b>	<b>C1</b>							G21	5.29	B2
S9	6.64	C1	W1	7.95	C2							G15	5.84	C1
S24	6.96	C1	W23	7.95	C2							G24	6.89	C1
S25	6.96	C1	W9	8.21	C2							G5	7.44	C1
S10	9.33	C2										G25	8.08	C1
S16	9.33	C2										G23	8.92	C2
S28	9.33	C2										G10	9.48	C2
S1	10.44	C2										G16	10.19	C2
S23	12.15	C2										G1	12.62	C2

### 2.2.3 Analysis of judgements from CEFR Table 3 and Manual 5.8

Two Familiarisation tasks followed a different format. Judges were given Table 3 from the CEFR and Table 5.8 from the Manual and were asked to reconstruct them by placing the descriptors in the correct cell. Appendix 4 (p. 85) presents the qualitative aspects of spoken language use from Table 3 (Council of Europe, 2001: 28-29) as given to the judges. The same layout was used for the written assessment criteria grid in Table 5.8 of the Manual (Council of Europe, 2003: 82). Judges were then given the descriptors already chopped up in a sealed envelope for each Table and were asked to stick them on it.

This task however poses a challenge for statistical analysis. First, apart from distinguishing between levels as in the previous tasks, judges also need to distinguish among different aspects of speaking and writing, such as range, coherence, accuracy, etc. Therefore, results should be treated with caution, because of the format of the task. This is because when a cell is filled with a wrong descriptor, this automatically results in a second wrong level assignment.

Following the above, analysis of judgements for these tasks will focus on raw scores, i.e. the number of correctly placed descriptors. Table 2.8 summarises these scores.

Table 2.8 Correct placement of descriptors for CEFR Table 3 and Manual Table 5.8

Descriptors	Descriptors	Mean	Min	Max	SD
Speaking CEFR Table 3	30	22.36	15	28	4.2

In general judges were successful as revealed by the mean score; however speaking resulted in a lower score. Interestingly, lower scorers, usually below 20, expressed their dislike of the task and stated a preference for the other format; this might suggest that the Familiarisation task format will indirectly affect the final score results and panellists who dislike a specific format might lose interest and motivation in completing the task. Further analysis could also concentrate on which categories of the two Tables appear to be hard to distinguish, but since the focus of the present analysis is the rater rather than the descriptors, such analysis is not included here.

### **2.3 Conclusion**

The analysis of judgements has clearly suggested that the Trinity panel is consistent and has a good understanding of the CEFR levels. According to the Manual, this is essential before proceeding to the next phases of the linking process. Statistical analysis of their judgements has revealed the following:

- High intra- and inter-rater reliability
- High rank order agreement between the CEFR and the judges
- Correct scaling by judges
- Positive effect of discussion; this was also pointed out by judges as very helpful to clarify any misunderstandings regarding the level of the descriptors.

To conclude, it would be reasonable to argue that the panel consists of expert judges who are expected to produce reliable decisions in the following two phases discussed below.

### 3 Specification

In this section the Specification phase of the project is described. I will first describe the methodology and then the results of this phase. The aim of the Specification was to build a linking claim about the content relevance of the exam to the CEFR, using the Forms provided by the Manual. However, before working with the Specification Forms, judges did a number of Familiarisation tasks in order to investigate whether their understanding of the CEFR scales was as consistent as in the Familiarisation phase. The results of the Familiarisation tasks will be discussed first, in subsection 3.1. The full programme of the Specification meeting that took place in November 2005 can be found in Appendix 2 (p. 83).

#### 3.1 Familiarisation tasks

The Manual strongly recommends that Familiarisation tasks should be organised at the beginning of the Specification phase and, aligned with that, judges were given the same tasks used in the Familiarisation meeting. The five descriptor sets from CEFR Table 2 descriptors, namely those for Speaking, Writing, Listening and Reading, plus the Global scale were given to the judges, asking them to identify the level (A1-C2). Only this task format was used this time for practical reasons, mainly time limitations, thus the tasks of reconstructing the cells of CEFR Table 3 and Manual Table 5.8 were not included.

The analysis of judgements is identical to the previous section; classical statistics and the Rasch model were used in order to establish judges' consistency. For this reason I will not repeat details on methodology, which replicated that in the previous section. The following subsections will therefore concentrate on summarising the results from the Familiarisation tasks that took place on the first day of the Specification meeting.

##### 3.1.1 Analysis of judgements using classical statistics

Classical statistics were employed to investigate intra and inter-rater reliability, internal consistency of the group and agreement with the CEFR. Results are summarised below.

Table 3.1 shows intra-rater reliability; this was calculated by running Spearman correlations of each judge's ratings with the ratings in the two rounds of the September meeting. Overall high levels of intra-rater reliability are observed, with the lowest value observed for Writing as was the case in the Familiarisation meeting.

*Table 3.1 Intra-rater reliability-summary statistics*

Scales	September 1 <sup>st</sup> round			September 2 <sup>nd</sup> round		
	Mean*	Min	Max	Mean*	Min	Max
Speaking	0.927	0.895	0.960	0.945	0.916	0.993
Writing	0.894	0.781	0.955	0.891	0.833	0.953
Listening	0.922	0.849	0.959	0.932	0.829	0.989
Reading	0.936	0.830	0.969	0.938	0.854	0.966
Global	0.943	0.892	0.983	0.939	0.886	0.966

\*Average using Fisher's Z-transformation

High levels of inter-rater reliability and internal consistency are observed in Table 3.2, with some moderate correlations for the minimum of Writing and Reading.

*Table 3.2 Inter-rater reliability and internal consistency-summary statistics*

Scales	Inter-rater reliability			Alpha
	Mean*	Min	Max	
Speaking	0.927	0.864	0.971	0.990
Writing	0.866	0.745	0.956	0.985
Listening	0.911	0.818	0.973	0.990
Reading	0.894	0.763	0.960	0.988
Global	0.918	0.852	0.960	0.990

\*Average using Fisher's Z-transformation

Table 3.3 summarises Spearman correlations run between each judge's levels and the correct CEFR levels in order to examine agreement of the judges' ratings and the correct rank order of the descriptors. Coefficient  $\kappa$  is also included in order to show exact agreement between panellists and the CEFR levels. As stated in subsection 2.2.1, this coefficient cannot be calculated when a judge has not used the full range of levels, which is the N size of 9 in the first three sets of descriptors. It appears that judges can understand the rank order of the descriptors as shown by Spearman, but since Kappa is relatively low, some adjacent levels are mixed up. In the first three descriptor sets one judge did not use the whole range of levels as can be seen in the last column; therefore  $\kappa$  could not be calculated for this person.

*Table 3.3 Rater-CEFR agreement-summary statistics*

Scales	Spearman correlations			Cohen's Kappa			N
	Mean*	Min	Max	Mean	Min	Max	
Speaking	0.953	0.912	0.990	0.61	0.32	0.92	9
Writing	0.898	0.818	0.965	0.5	0.286	0.802	9
Listening	0.938	0.849	0.976	0.56	0.235	0.803	9
Reading	0.932	0.894	0.964	0.52	0.345	0.697	10
Global	0.949	0.887	0.983	0.63	0.361	0.88	10

\*Average using Fisher's Z-transformation

Table 3.4 presents descriptive statistics of the scores obtained in the tasks, that is, the number of descriptors placed at the correct level for each of the 5 sets. Scores are in general higher than the first round of the September meeting, again with the exception of slightly lower scores for Reading.

*Table 3.4 Scores obtained by judges in the familiarisation tasks-correct answers*

Scales	Descriptors	Mean	Min	Max	SD
Speaking	30	19.4	12	28	5.13
Writing	25	14.3	9	21	3.77
Listening	19	11.7	7	16	3.2
Reading	20	12.1	9	15	1.85
Global	30	20.7	14	27	4.83

Overall, classical statistics show comparable results to the Familiarisation phase. High levels of consistency are observed, but again with some reservations as to whether judges can distinguish between adjacent levels as  $\kappa$  coefficient and raw scores show. As already stated in the previous section, this is expected due to the breaking down of the original descriptors into smaller statements.

### 3.1.2 Analysis of judgements using the Rasch model

Employing the Rasch model using the FACETS programme generated some interesting results. Table 3.5 summarises results for the rater facet.

*Table 3.5 Judges' characteristics- logits and infit mean square values*

Scales	logits (min-max)	Infit mean square (min-max)	Reliability
Speaking	-1.33-.65	-.33-1.18	.52
Writing	-3.40 (-1.48)-(1.36) 2.01	.42-1.05	.90
Listening	-3.12 (-1.48)-(1.89) 2.02	.51-(1.18) 2.25	.84
Reading	-1.00-1.10	.46-1.54	.36
Global	-.95-.71	.45-1.25	.44

The high reliability of Writing and Listening indicates that there are differences in rater leniency/harshness. This is confirmed by the logits column, where values in parentheses also show the next closest value. Closer investigation of the FACETS output file and the EXCEL spreadsheet sheds some light on the reasons for the observed differences.

For Writing in particular, Matt appears to be the most lenient because he assigned A2 and B1 when the others assigned higher levels, thus he appears lower on the rater scale. Lora is the strictest, because she was the only one to assign A2 and B1 to descriptors that all other judges rated as B1 and B2 respectively. The logit values of the other judges, as shown in the parentheses, indicate that the spread of the judges on the rater column is narrower.

For Listening, the values of 2.02 and -3.12 logits belong to Lora and Nicola respectively. Comparison of their ratings suggests that in six cases out of eight, Lora assigned a lower level to the descriptors, therefore she appears as stricter. It should be stressed that results might have also been affected by the fact that four Listening descriptors, 3 at A1 and 1 at C2 were excluded by FACETS for the estimation of rater leniency/severity. This is because these descriptors, which generated absolute agreement among raters as to their level, do not provide any information as to rater behaviour. In fact, Bond and Fox (2001:13) recommend deletion of such items, since they do not say much about the person facet. With 15 descriptors remaining, disagreement in more than half of them might have contributed to the difference of these two raters in leniency/severity. All these, of course, suggest that a plain two-facet model needs to be carefully analysed in this particular context in order to generate accurate conclusions regarding rater behaviour. This is because a rater's leniency or severity is established in relation to the other raters and not the correct level of the descriptors. Finally, the infit value of 2.25 comes from a judge who is probably flagged as inconsistent because she was the only one to assign C2 to a B2 descriptor, which was estimated by the other judges as either B1 or B2.

To summarise the above, even though results in Table 3.5 might not be very impressive for Listening and Writing, the reasons for extreme values have been explained here, indicating that the group discussion after rating each descriptor set is crucial to help judge realise any inconsistencies in their ratings, as is the case for example with the misfitting rater in Listening.

A very important aspect that was also examined in the previous section, is how the descriptors are scaled from lower to higher, according to the judgements of the panel. Table 3.6 illustrates the scaling for all descriptor sets, with the highlighted descriptors flagging incorrect scaling. Scaling appears to be very close to the expected pattern. It should be stressed, however, that the number of highlighted descriptors is closer to the first round of the Familiarisation meeting, rather than the second. In turn this shows that even if a detailed Familiarisation phase has been organised, as was the case with the September meeting of this project, Familiarisation is needed just before Specification and should be accompanied by group discussion; some false understanding of the CEFR scales might appear even after a detailed Familiarisation and group discussion should focus on ensuring this is not the case before Specification.

Table 3.6 Scaling of the descriptors

Speaking			Writing			Listening			Reading 1			Global		
Descr.	Logit value	CEFR level	Descr.	Logit value	CEFR level	Descr.	Logit value	CEFR level	Descr.	Logit value	CEFR level	Descr.	Logit value	CEFR level
S3	-11.81	A1	W24	-12.29	A1	L3	-12.59	A1	R6	-9.58	A1	G12	-11.20	A1
S13	-10.90	A1	W25	-11.54	A1	L4	-12.59	A1	R19	-9.58	A1	G13	-10.18	A1
S20	-10.90	A1	W6	-10.87	A2	L6	-12.59	A1	R4	-6.93	A1	<b>G11</b>	<b>-7.97</b>	<b>A2</b>
S12	-9.78	A1	W19	-10.87	A2	L12	-9.50	A2	R12	-6.03	A2	G3	-7.51	A1
<b>S22</b>	<b>-9.32</b>	<b>A2</b>	W4	-7.16	A2	L15	-9.50	A2	R16	-5.65	A2	G20	-7.51	A1
<b>S30</b>	<b>-9.32</b>	<b>A1</b>	W12	-4.60	B1	L16	-9.50	A2	R15	-4.63	A2	<b>G29</b>	<b>-7.51</b>	<b>A2</b>
S4	-7.90	A2	W16	-4.60	B1	L11	-3.50	B1	R7	-3.68	B1	<b>G30</b>	<b>-7.51</b>	<b>A1</b>
S29	-7.90	A2	W15	-1.95	B1	L5	-0.75	B1	R11	-3.68	B1	G4	-7.05	A2
S17	-7.34	A2	W11	-1.29	B1	L7	-0.21	B1	R8	-3.00	B1	G17	-6.57	A2
S2	-5.84	B1	W8	.91	B2	L8	.83	B2	R17	-2.64	B1	G22	-6.09	A2
S19	-5.84	B1	W7	3.98	B2	L2	2.61	B2	R5	-2.28	B1	G19	-3.30	B1
<b>S11</b>	<b>-4.91</b>	<b>A2</b>	W22	3.98	B2	L13	3.01	B2	R13	2.03	B2	G26	-3.30	B1
S8	-2.12	B1	W5	4.48	B2	L17	4.18	B2	R2	2.26	B2	G2	-2.39	B1
S27	-2.12	B1	W17	5.32	C1	<b>L18</b>	<b>4.55</b>	<b>C1</b>	<b>R18</b>	<b>2.71</b>	<b>C1</b>	G27	-1.96	B1
S26	-1.44	B1	W13	6.89	C1	<b>L19</b>	<b>4.55</b>	<b>B2</b>	<b>R20</b>	<b>2.94</b>	<b>B2</b>	G18	-1.54	B1
S18	-.80	B1	<b>W9</b>	<b>7.16</b>	<b>C2</b>	L9	4.92	C1	R9	3.17	C1	G8	-.74	B1
S14	1.84	B2	<b>W14</b>	<b>7.16</b>	<b>C2</b>	L10	5.66	C1	R10	3.65	C1	G14	1.94	B2
S6	2.64	B2	W2	7.44	C1	L14	8.27	C1	R1	3.91	C1	G28	2.39	B2
S21	3.05	B2	W21	7.44	C1	L1	10.12	C2	R14	5.11	C1	G6	4.18	B2
S15	4.28	C1	W1	7.71	C2				R3	7.95	C2	<b>G9</b>	<b>4.18</b>	<b>C1</b>
S5	4.68	C1	W3	7.71	C2							G21	4.18	B2
<b>S7</b>	<b>5.08</b>	<b>B2</b>	W10	7.71	C2							G7	4.54	B2
S25	5.08	C1	W18	7.99	C2							G15	4.87	C1
S9	5.48	C1	W23	7.99	C2							G5	5.46	C1
S24	6.63	C1	<b>W20</b>	<b>8.28</b>	<b>C1</b>							G24	6.27	C1
S10	8.21	C2										G25	7.08	C1
S16	8.21	C2										G10	7.72	C2
S1	10.70	C2										G23	8.13	C2
S28	10.70	C2										G16	8.68	C2
S23	12.04	C2										G1	9.52	C2

### 3.1.3 Conclusion

To conclude, despite some discrepancies observed, consistency of judgements is high, as in the Familiarisation meeting in September 2005. Judges also felt that the group discussion was very useful in pointing out any incorrect levels placement for

the descriptors; all judges then unanimously expressed their confidence in moving to the completion of the Specification Forms.

## **3.2 Methodology**

The methodology of the Specification phase is described here in chronological order in three parts.

### **3.2.1 Before the meeting**

The group of judges was the same as in the Familiarisation meeting. However, this time 2 judges could not come and the group comprised 10 people. This phase aims according to the Manual at describing the content of the exam to be linked to the CEFR. A number of Forms in the Manual (Ch. 4) should be filled in, providing general and detailed description of the exam in relation to the CEFR.

General description Forms A1-A7 were sent to Trinity approximately 3 weeks before the meeting. Three ESOL managers worked together to fill them in and they were returned prior to the Specification meeting and discussed with the project coordinator. These Forms were not given to the panel, as they contained administrative details which the panel would not probably know. Additionally, it was felt that three people would be enough to complete the Forms and therefore, there was no need to add to the judges' workload during the three-day meeting. The Forms were used twice, once for GESE and once for ISE.

Based on the test specifications and judges' recommendations by email prior to the meeting, the following detailed description Forms were chosen: For GESE: A8 (Overall Impression), A11 and A13 (Spoken Interaction and Production), A9 (Listening Comprehension) and A19-A21 (Aspects of Competence in Reception, Interaction and Production respectively). For ISE: All Forms used in the mapping of GESE, plus A12 and A14 (Written Interaction and Production), A10 (Reading Comprehension) and A15-A16 (Integrated Skills)

For GESE, even though Listening is only assessed explicitly at the higher levels, judges felt that it should also be included in the lower grades, because it is part of Spoken Interaction and is included in the marking criteria as such. Interestingly, the judges did not express the need for Forms A15-16 to be completed, mainly because there is no combination of Listening and Spoken Interaction in them. Despite the fact that the Manual encourages its users to add any aspects not covered in the Forms, the judges felt that the Integrated Skills Form of the Manual would not add any information for mapping GESE. However, Forms A15 and A16 were unanimously included in the mapping of ISE, since the test specifications mention explicitly the integrative approach of the suite in the Controlled Written component (Reading into Writing).

Apart from the Manual Forms, further documentation used during the Specification meeting included the syllabus for each suite (Trinity College London, 2005a, 2005b) and the CEFR volume. Additionally, all relevant scales from the CEFR and lists from *Threshold 1990* (van Ek & Trim, 1998) were included in a handout in order to help judges find easily the scales and lists they would need to complete the Forms.

Given that judges would need to fill in the Specification Forms of the Manual for a total of 16 levels (12 GESE Grades and 4 ISE levels) during the three days of the meeting, time limitations were considered; Specification was organised in parallel sessions using three teams as can be seen in Table 3.7. Four teams were used for ISE. Since 10 judges took part, the GESE sessions comprised groups of three or four

members, whereas the ISE groups had two or three members. After each session, a plenary would follow during which each team would point out any problems during the completion of the forms.

*Table 3.7 Organisation of the parallel sessions for GESE and ISE*

<b>GESE Initial Stage</b>	<b>GESE Elementary Stage</b>	<b>GESE Intermediate Stage</b>	<b>GESE Advanced Stage</b>	<b>Extra session for each GESE stage</b>	<b>ISE</b>
Grades 1-3	Grades 4-6	Grades 7-9	Grades 10-12	4 GESE stages	Levels 0-III
Team 1-Grade 3	Team 1-Grade 6	Team 1-Grade 8	Team 1-Grade 10	Team 1-Initial	Team 1-ISE I
Team 2-Grade 1	Team 2-Grade 4	Team 2-Grade 7	Team 2-Grade 11	Team 4-Elementary	Team 2-ISE 0
Team 3-Grade 2	Team 3-Grade 5	Team 3-Grade 9	Team 3-Grade 12	Team 3-Intermediate	Team 3-ISE III
				Team 2-Advanced	Team 4-ISE II

### **3.2.2 During the meeting**

As pointed out earlier, the Manual states that Familiarisation tasks should be organised before any attempt to map the content of the tests on the CEFR. Therefore, judges were asked to repeat the tasks they did in September. These are the tasks analysed in subsection 3.1. Discussion regarding the CEFR descriptors followed, immediately after each descriptor set, and panellists were invited to consider their judgements which were shown on a data projector and explain how they arrived at decisions regarding level placement. Participants were also given feedback on their judgements during the September meeting.

The aims of the Specification and the documentation that would be used were then presented, following judges' assertion that they felt confident to proceed to this phase after the Familiarisation. The programme on page 83, which was given to judges before the meeting, was followed closely with the exception of the first day: the teams needed additional time for the Initial Grades as they were not familiar with the Forms and the panel had to agree on shared understanding of some of the questions in the Forms. A lot of discussion had to do with the Specification of tasks and communicative tasks in the Detailed Description Forms. It was decided that tasks as a general term would correspond to the components of the Grades, for example Topic presentation, Topic discussion, etc, (see also subsection 1.4 of this Report). Communicative tasks would correspond to what the syllabus describes as communicative skills. Finally, communicative activities would refer to language functions described in the syllabus.

The additional GESE session was decided on because the plethora of GESE Grades would result in several placements of Grades at the same CEFR level. After finishing with ISE, the panel spent the afternoon investigating further the GESE grades. Four teams were again formed, 1 per GESE stage (see Table 3.7, column 5). The teams were asked to refer to the documentation mentioned above and state the differences of Grades placed at the same level. The discussion that followed finalised the mapping of the GESE Grades on the CEFR.

### 3.2.3 After the meeting

At the end of each session, team members provided the project coordinator with a final set of Forms describing the Grade they mapped on the CEFR, which were later word-processed. The panel was asked to fill in a feedback form, and if possible to reply to questions regarding what they liked/did not like during the meetings, what they saw as beneficial for Trinity during this process and what recommendations they could make regarding the use of the CEFR and the Manual.

To avoid needless repetition when reporting the results of the phase, results were per GESE stage rather than per Grade when description of the test content was required in the Manual Forms (e.g. “What tasks are the test-takers expected to be able to handle”). Results per Grade were reported when CEFR levels should be indicated, which is the last question in each Manual Form. The decision to group results per stage was based on the expectation that the description of the content of the test would be similar since the format is exactly the same for the Grades of each stage. For example, Grades at the Elementary stage contain two tasks which are similar across the stage and answering the same questions for the Grades of the Elementary stage would be mere repetition. The completed Specification Forms are included in a separate supplement to the present report.

Judgements were made in a sequence of rounds as can be seen in Table 3.8. It was expected that this kind of judgement iteration would provide the necessary agreement regarding the results to be reported to the Council of Europe and would adequately reflect the Trinity suites in relation to the CEFR. It would also ensure that judgements made by the coordinator when word-processing the Forms would not dominate those by the panel or the managers and vice-versa.

*Table 3.8 Rounds of judgements during the Specification stage*

<b>General description</b>	<b>Detailed description</b>
Round 1: Forms filled in by Trinity ESOL managers	Round 1: Selection of Forms after discussion on the mail list among the coordinator and the panel members
Round 2: Completed Forms checked by the project coordinator who returned them with feedback to the managers	Round 2: Completion of Forms during the Specification meeting-group work.
Round 3: Managers finalising Forms' completion based on suggestions by the coordinator	Round 3: Discussion of decisions for each GESE/ISE stage- plenary discussion during the Specification meeting. Round 4: Justification of the levels assigned in Round 2 by considering Grades belonging to the same stage-group work for GESE only, examining adjacent Grades that were placed at the same CEFR level Round 5: Group discussion finalising the levels for both GESE and ISE- plenary discussion. Round 6: Consideration of the results by the project coordinator. Judgements were analysed and examined closely in relation to the CEFR and test specifications. Judgements standardised and reported in word-processed format. Round 7: Word-processed Forms posted to each judge. Judges are asked to check the answers and provide feedback, especially if they do not agree with the way results are reported. Round 8: Report of the Specification results is finalised by taking under consideration feedback from judges.

### 3.3 Results

The results of the analysis of the Manual Forms are graphically presented below. These graphic profiles are basically an adaptation of Form A23 of the Manual





Table 3.9 Relationship of GESE content to the CEFR-holistic estimation

Stage	Grades	CEFR level
Initial	Grade 1	A1
	Grade 2	A1/A2
	Grade 3	A2.1
Elementary	Grade 4	A2.2
	Grade 5	B1.1
	Grade 6	B1.2
Intermediate	Grade 7	B2.1
	Grade 8	B2.1/B2.2
	Grade 9	B2.1/B2.2
Advanced	Grade 10	B2.2/C1
	Grade 11	C1
	Grade 12	C1/C2

Table 3.10 Relationship of ISE content to the CEFR-holistic estimation

Stage	CEFR Level
ISE 0	A2/A2.2/B1.1
ISE I	B1.1/B1.2
ISE II	B2.1/B2.2
ISE III	C1

### 3.4 Discussion of methodology and results

The aforementioned results and the group discussion during the Specification phase raise a number of issues regarding the methodology and the results and will be dealt with in this section.

#### 3.4.1 The branching approach

The CEFR suggests a branching approach (Council of Europe, 2001:31-33) where additional levels are defined apart from the six ones. This was very suitable for the GESE suite, because of the 12 Grades. However, only A2, B1 and B2 levels have an additional '+' level description, and this is not the case with all CEFR scales. This is the reason why GESE Grades appear to cluster around the same level in the Figures above; it would be very superficial to place Grades on additional levels for A1 and C1 for example when they don't exist in the CEFR.

#### 3.4.2 Fit between the CEFR descriptors and the Trinity suite

The judges pointed out several times during the meeting that the CEFR descriptors did not always reflect the content of Trinity exams. For example A1 level in the Global scale states that language learners "can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has" (Council of Europe, 2001:24). Judges noted that Grades 1-3 do not have any questions on behalf of the candidate, but apart from the function of asking questions A1 descriptors seem to describe adequately the Trinity context. The judges also noted that part of a descriptor would not apply at all, without however meaning that the

candidates are unable to do what the CEFR describes. For that reason, all answers in the Specification Forms regarding the CEFR level of GESE Grade or ISE levels state exactly which part of a descriptor describes Trinity candidates' performance, by quoting the exact descriptor unit. Finally, as the graphic profiles show, judges usually find appropriate description for GESE Grades and ISE levels in more than one CEFR level. This perhaps is an indication that performance by candidates receiving higher scores at a Grade might have characteristics usually at the '+' CEFR levels, thus the definition of the borderline candidate, that is the candidate who just passes a GESE Grade or ISE level, is of particular importance. The importance of defining the performance of the borderline candidate is the focus of Standardisation.

### **3.4.3 Justification and rationale behind decisions**

Following the issue of fit between descriptors and the exam, the project coordinator pointed out several times in the Specification sessions the importance of making substantive judgements and providing a clear rationale for decision making. As can be seen in the Specification Forms, the amount of justification is notable and the rationale behind decisions is explained as clearly as possible. This has hopefully contributed to the validity of the Specification claim.

### **3.4.4 The link between GESE and ISE**

The ISE syllabus claims that ISE levels are "linked to Trinity's Graded Examinations in Spoken English, and at each of the four chosen levels, (grades 4, 6, 8, 11), they represent an integrated skills alternative to the grades in spoken English" (Trinity College London, 2005b:7). Graphic profiles of the exams suggest that the judges found this to be the case after describing test content using the CEFR, even though more than one CEFR level was observed for the two lower ISE levels.

### **3.4.5 The skill of Listening**

The CEFR (Council of Europe, 2001:24) claims that the candidate can "understand virtually everything heard". The judges felt that such an expectation is even beyond the abilities of many native speakers and decided that the listening level of Grade 12 should be C1, because of the above statement that belongs to C2.

### **3.4.6 Using the Specification Forms in practice**

Negative comments were made regarding use of the Forms of the Manual. The judges found them very repetitive and that was clear from the tendency in the panel to refer back to a previous Form when a new one would ask the same question. For example Forms A19, A20 and A21 ask more or less the same questions, and after completing A19, judges would just mention "see Form 19" in A20 and A21. This was very carefully examined here, as it should be ensured that aspects of linguistic competence for example are the same in Production, as well as Interaction and Reception. Moreover, it was evident during the Specification stage that the workload was very heavy, as the judges had to work intensively. Wrong or incomplete reference to the CEFR was evident in the Forms and was carefully examined so that the most accurate information is included in the supplement to the present report.

### **3.4.7 Validity of the Specification claim**

The Manual is not very helpful on how to organise and run the Specification phase and how to analyse results. As I have already argued recently (Papageorgiou, 2006), building a valid Specification claim is rather complicated, because group dynamics play an important role in decision making. Following this, the panellists were audio recorded while working on completing the Specification Forms and the transcribed discussions were analysed for the purposes of the author's PhD using a social psychology framework which has recently been used in language testing for investigating a group task of compiling test specifications (Davidson & Lynch, 2002). This part of the thesis-in-progress investigates how this framework can be applied to the linking context and how it can shed light on aspects that affect the validity of the linking claim during Specification. I will only briefly summarise here some of the points discussed in the thesis chapter, which can be later consulted by any interested readers.

A claim in favour of the validity of the Specification linkage could be located in the fact that the small groups exhibited high levels of what the literature calls "cohesiveness" (Davidson & Lynch, 2002:101). Cohesiveness has been defined in various ways but generally refers to how well group members get along with each other, how positively they identify with the group and how much they want to remain in the group. Cohesiveness can be assessed by asking group members how much assistance they receive, by investigating the group's perceptions of the friendliness and cooperation within the group and by examining the degree to which there is contact between the group members outside of group meetings. Davidson and Lynch (2002: 102) suggest that "we need to foster an atmosphere of support and collaboration which is focused on a clearly understood commitment to the group's task", which has a clear implication on the atmosphere which should be created when the task is to work on the Specification stage. Apart from observing that this was the case during the sessions, feedback forms given to judges contained a question as to whether they enjoyed working with the other group members. All judges responded positively.

It was also found that another aspect could threaten the validity of the Specification linkage, which was traced during the sessions and was treated accordingly. This aspect has to do with the amount of ambiguity to be found during the decision making process (Davidson & Lynch, 2002; Salazar, 1995). As I report in my thesis-in-progress, terminology problems and lack of definition in the CEFR pointed out by Alderson et al. (2006) contribute to the amount of ambiguity during the decision making process. This underlines the importance of the role of the project coordinator during the Specification; as expected, the judges were in need of clarification of terminology used in the CEFR and the project coordinator had to be prepared to assist, thus minimising the amount of ambiguity during decision making. This was evident during the first time judges worked with the CEFR for describing the content of the Initial Grades, which, as I mentioned earlier, lasted longer than originally planned.

Whilst most aspects of group dynamics discussed in Davidson and Lynch (2002) such as roles, power and status did not appear to have an effect on decision making, there was one aspect that could be considered as a counterclaim of the validity of the Specification linkage. As I have pointed out elsewhere, (Papageorgiou, 2006) working with judges who have a relationship with the exam provider means that these judges are aware of the expectations as to the level of the test and inevitably bring some bias. This is evident when, for example, judges instead of looking at a range of levels for

the test, only attempt to see whether descriptors of the intended level describe the test content adequately. In order to avoid that, the project coordinator stressed at the beginning of the Specification stage that judges should consider a range of levels and provide a clear rationale behind their decision. As suggested above, the Forms supplement contains a good amount of justification, but the extremely consistent pattern of judgements in the graphic profiles above might also suggest that the inevitable bias of insiders was not totally avoided and that the judges were also influenced by the level of the corresponding GESE Grades described in the syllabus.

### **3.5 Conclusion**

This section has provided a description of the methodology and the results of the Specification phase. The outcome of this phase is a claim as to the content relevance of the exam to the CEFR summarised in subsection 3.3. The completed Detailed and General Description Forms in the supplement provide a wealth of information regarding the content analysis of the exam and can be consulted by test users who wish to learn more about the exam content in relation to the CEFR.

The Specification linkage will be further examined in this report in the next section, namely Standardisation.

## 4 Standardisation

In this section I report on the Standardisation phase (Council of Europe, 2003:Ch. 6), which took place at the Trinity Head Office in London, between the 28<sup>th</sup> of February and 2<sup>nd</sup> of March 2006. Appendix 3 (p. 84) contains the programme of the meeting. Eleven panellists participated, since one of the original twelve could not attend.

Following the methodology of the Manual, I will discuss the three phases of Standardisation, namely Training in subsection 4.3, Benchmarking in subsection 4.4 and Standard Setting in subsection 4.5. The methodology of this phase is described below, in subsection 4.2.

The aim of the Standardisation is to further examine the link between the exam and the CEFR by considering actual candidate performance and finally to establish cutscores in relation to the CEFR.

### 4.1 Familiarisation tasks

The Manual strongly recommends that Familiarisation tasks should be organised at the beginning of the Standardisation phase and, as was the case with Specification, judges repeated the tasks of the Familiarisation meeting. CEFR Table 2 descriptors, namely those for Speaking, Writing, Listening and Reading, plus the Global scale were given to judges, asking them to identify the level (A1-C2).

The analysis of judgements is identical to the previous sections, including classical statistics and the Rasch model which were used in order to establish judges' consistency. Results are discussed below, whereas details on the methodology can be found in subsection 2.1 (p. 12).

#### 4.1.1 Analysis of judgements using classical statistics

Classical statistics were employed to investigate intra and inter-rater reliability, internal consistency of the group and agreement with the CEFR. Results are summarised below.

Table 4.1 shows intra-rater reliability by running Spearman correlations of each judge's ratings with the ratings in the two rounds of the Familiarisation meeting in September 2005 and the Specification meeting in November 2006. Overall, high levels of intra-rater reliability are observed with only some moderate minimum values for Writing and Listening in the third column of the Table.

*Table 4.1 Intra-rater reliability-summary statistics*

Scales	September 1 <sup>st</sup> round			September 2 <sup>nd</sup> round			November		
	Mean*	Min	Max	Mean*	Min	Max	Mean*	Min	Max
Speaking	0.902	0.832	0.958	0.929	0.870	0.972	0.943	0.878	0.980
Writing	0.912	0.760	1	0.898	0.837	0.955	0.917	0.839	0.976
Listening	0.904	0.780	0.947	0.907	0.807	0.960	0.920	0.825	0.960
Reading	0.960	0.916	1	0.962	0.918	1	0.953	0.920	0.970
Global	0.948	0.896	0.979	0.953	0.866	1	0.950	0.909	0.983

\*Average using Fisher's Z-transformation

High levels of inter-rater reliability and internal consistency are also observed in Table 4.2; again, there are moderate correlations for the minimum of Writing and Reading.

*Table 4.2 Inter-rater reliability and internal consistency-summary statistics*

Scales	Inter-rater reliability			Alpha
	Mean*	Min	Max	
Speaking	0.924	0.847	0.975	0.990
Writing	0.868	0.718	0.955	0.986
Listening	0.883	0.722	0.953	0.988
Reading	0.948	0.888	0.991	0.994
Global	0.940	0.859	0.974	0.988

\*Average using Fisher's Z-transformation

Table 4.3 summarises Spearman correlations run between each judge's levels and the correct CEFR levels in order to examine agreement of the judges' ratings and the correct rank order of the descriptors. Coefficient  $\kappa$  is also included in order to show exact agreement between panellists and the CEFR levels. As already pointed out in previous sections, it appears that judges can understand the rank order of the descriptors as shown by Spearman, but because  $\kappa$  is relatively low, some adjacent levels are mixed up. One judge did not use the whole range of levels for Writing, nor did three judges for Listening as can be seen in the last column; therefore  $\kappa$  could not be calculated in these cases.

*Table 4.3 Rater-CEFR agreement-summary statistics*

Scales	Spearman correlations			Cohen's Kappa			N
	Mean*	Min	Max	Mean	Min	Max	
Speaking	0.946	0.887	0.981	0.553	0.28	0.839	11
Writing	0.913	0.777	0.956	0.536	0.333	0.805	10
Listening	0.928	0.818	1	0.634	0.369	1	8
Reading	0.969	0.925	0.994	0.713	0.461	0.938	11
Global	0.962	0.874	0.991	0.691	0.319	0.92	11

\*Average using Fisher's Z-transformation

*Table 4.4 Scores obtained by judges in the familiarisation tasks-correct answers*

Scales	Descriptors	Mean	Min	Max	SD
Speaking	30	18.82	12	26	4.49
Writing	25	14.36	4	21	4.74
Listening	19	12.18	6	19	3.66
Reading	20	15.18	11	19	2.79
Global	30	22.27	13	28	4.98

Table 4.4 presents descriptive statistics of the scores obtained in the tasks, that is, the number of descriptors placed at the correct level for each of the 5 sets. Scores are comparable to those in previous phases. The only exceptions are the very low minimum scores for Writing and Listening, obtained by the same person. This judge

did not also use the whole range of the CEFR levels as was pointed out above when discussing results of the  $\kappa$  coefficient. As I will explain later, such a poor performance is of concern because it might mean that standards set by this person for Writing and Listening are not very reliable; thus, excluding her judgements from the final cut scores in relation to the CEFR will be considered.

Overall, classical statistics show comparable results to the Familiarisation and Specification phases. High levels of consistency are observed, but again with some reservations as to whether judges can distinguish between adjacent levels as  $\kappa$  coefficient and raw scores show. Even though this might be expected due to breaking down of the original descriptors into smaller statements, exclusion of one judge from the final cut scores for Writing and Listening will be considered in subsection 4.5.3.

#### 4.1.2 Analysis of judgements using the Rasch model

Results of FACETS analysis are discussed here. Table 4.5 summarises results for the rater facet. Numbers in parentheses indicate the infit mean square value that is the next closest to the highest value observed. The high reliability for four sets of descriptors suggests that judges differ in terms of leniency/harshness; possible reasons are given below.

*Table 4.5 Judges' characteristics- logits and infit mean square values*

Scales	logits (min-max)	Infit mean square (min-max)	Reliability
Speaking	-1.10-1.70	.43-(1.12) 1.60	.83
Writing	-1.15-1.27	.43-(1.39) 1.77	.70
Listening	-1.25-1.26	.31-(1.53) 1.81	.47
Reading	-2.50-3.67	.42-(1.37) 1.45	.82
Global	-1.63-2.90	.48-(1.31) 1.88	.84

For Speaking, three judges, namely Tim, George and Rita had values of -1.10, 1.05 and -1.03 respectively and appear to be the most lenient, whereas Claudia, the judge with the low scores for Writing and Listening appears higher on the scale with a value of 1.70 and thus she is stricter in relation to other judges. Comparison of scores among these raters shows that Claudia tends to assign lower levels. The logit values range excluding these four raters is between -.79 to 1.24. Rita is also the one with infit mean square value of 1.60. She differs from other raters because she was the only one to assign C1 to a B2 descriptor when everybody assigned B1 and B2. She also chose B1 for an A2 descriptor when the others only assigned A1 and A2.

Rita also appears to be more lenient than others when rating Writing descriptors, with a logit value of -1.27. Matt is the strictest judge, at 1.15 logits. The logit values range excluding these raters is between -1.06 to .85. Claudia, the low scorer in Writing and Listening, is again different from other raters with an infit mean square of 1.77, whereas Erika is also slightly misfitting based on the 1.3 threshold, with 1.39. Investigation of their ratings in relation to those by other judges reveals that this could be because they assigned B1 and B2 to two C2 descriptors when other judges assigned C1 and C2.

The case of Reading is of special interest because one would expect to get good results, since the panel did very well as the raw scores show in Table 4.4. The top scorer, Sally, placed 19 out of 20 at the correct level, but she appears to be the most lenient rater with a logit value of -3.67. Another panellist, Andrew, with a score of 15 appears to be at the top of the rater scale with a tendency to assign higher levels and

thus being the strictest rater. This paradox could also have happened because FACETS has excluded from the analysis two A1 and one C2 descriptors, because all judges agreed on the level. As discussed in subsection 3.1.2, FACETS excludes such extreme observations. The logit values range excluding these raters is between -1.22 and 1.58. All the above remind us that the way the two-facet Rasch model has been applied here is primarily a way to make inferences about a rater's performance in relation to other raters, not necessarily in relation to the correct levels. For this reason, it is of interest to see why Lora is misfitting with an infit mean square of 1.45: going back to the data, this appears to be because she was the only one who placed a B1 descriptor at B2.

To summarise FACETS results, results in Table 4.5 reveal behaviour by some raters that needs further consideration. Remedy was again in the form of group discussion. After rating each descriptor, ratings were displayed on the screen and, as judges reported, this helped them realise any inconsistencies in their ratings.

The scaling of the descriptors is also examined here. Table 4.6 illustrates the scaling for all descriptors sets, with the highlighted descriptors flagging incorrect scaling.

Table 4.6 Scaling of the descriptors

Speaking			Writing			Listening			Reading 1			Global		
Descr.	Logit value	CEFR level	Descr.	Logit value	CEFR level	Descr.	Logit value	CEFR level	Descr.	Logit value	CEFR level	Descr.	Logit value	CEFR level
S23	10.79	C2	W18	7.88	C2	L1	10.42	C2	R3	12.54	C2	G1	12.39	C2
S28	9.24	C2	W1	7.62	C2	L14	6.40	C1	R14	9.83	C1	G10	10.24	C2
S16	8.33	C2	W10	7.40	C2	L9	4.96	C1	R1	8.95	C1	G16	9.66	C2
S1	8.33	C2	W23	7.40	C2	L10	4.04	C1	R10	8.95	C1	G23	9.19	C2
S10	6.79	C2	W3	7.19	C2	L18	3.74	C1	R18	7.04	C1	G25	8.39	C1
S25	4.59	C1	W14	6.99	C2	L17	3.13	B2	R9	6.95	C1	G24	7.69	C1
S24	4.59	C1	W2	6.62	C1	L19	3.13	B2	R13	6.34	B2	<b>G6</b>	<b>5.96</b>	<b>B2</b>
S15	3.86	C1	W13	6.62	C1	L2	2.49	B2	R20	6.34	B2	<b>G5</b>	<b>5.96</b>	<b>C1</b>
S9	3.86	C1	W20	6.43	C1	L13	2.15	B2	R2	5.26	B2	<b>G21</b>	<b>5.39</b>	<b>B2</b>
S7	3.19	B2	W21	6.43	C1	L8	1.45	B2	R5	-1.80	B1	<b>G15</b>	<b>5.39</b>	<b>C1</b>
S21	2.30	B2	<b>W9</b>	<b>6.24</b>	<b>C2</b>	L7	.31	B2	R8	-1.80	B1	<b>G9</b>	<b>5.09</b>	<b>C1</b>
S5	2.30	C1	W17	5.15	C1	L5	-.09	B1	R17	-1.80	B1	G7	4.41	B2
S6	1.44	B2	W22	4.50	B2	L11	-3.39	B1	R7	-2.93	B1	G28	4.03	B2
S14	.11	B2	W5	4.10	B2	L16	-7.05	A2	R11	-5.31	B1	G14	3.63	B2
S18	-1.11	B1	W7	3.10	B2	L15	-7.46	A2	R15	-10.95	A2	G8	.05	B1
S8	-1.11	B1	W8	-.20	B2	L12	-8.73	A2	R12	-15.69	A2	G18	-1.16	B1
S27	-1.57	B1	W11	-2.14	B1	L4	-10.34	A1	<b>R4</b>	<b>-16.82</b>	<b>A1</b>	G27	-1.83	B1
S26	-2.05	B1	W15	-2.14	B1	L6	-10.34	A1	<b>R16</b>	<b>-16.82</b>	<b>A2</b>	G2	-3.03	B1
<b>S17</b>	<b>-4.59</b>	<b>A2</b>	W12	-5.01	B1	L3	-12.61	A1	R6	-19.44	A1	G26	-3.55	B1
<b>S11</b>	<b>-4.99</b>	<b>A2</b>	W16	-6.21	B1				R19	-19.44	A1	G19	-4.55	B1
<b>S19</b>	<b>-5.41</b>	<b>B1</b>	W4	-8.70	A2							G22	-8.03	A2
<b>S2</b>	<b>-5.41</b>	<b>B1</b>	W6	-10.68	A2							G11	-8.03	A2
S29	-6.74	A2	W19	-10.68	A2							G29	-10.85	A2
S22	-7.65	A2	W24	-13.31	A1							G17	-11.42	A2
S4	-7.65	A2	W25	-14.73	A1							G4	-11.42	A2
S12	-8.56	A1										G3	-11.94	A1
S20	-9.53	A1										G20	-12.98	A1
S30	-10.09	A1										G30	-13.63	A1
S13	-10.09	A1										G13	-15.92	A1
S3	-10.78	A1										G12	-15.92	A1

Scaling appears to be satisfactory, with only a few mixed descriptors, which again stresses that despite having organised Familiarisation tasks in September and November, separate Familiarisation is required before Standardisation and should be accompanied by group discussion to point out any misplaced descriptors. In this Familiarisation session, judges were also shown the ruler map of the Global descriptors in order to receive additional feedback on their ratings.

### **4.1.3 Conclusion**

Similarly to Specification, despite some discrepancies observed, judges felt that the group discussion was very useful in pointing out incorrect level placement for the descriptors; all judges then unanimously expressed their confidence in moving to the next stages of the Standardisation phase.

## **4.2 Methodology**

The methodology of this phase is described in this subsection in chronological order.

### **4.2.1 Before the meeting**

The preparation for the meeting started immediately after the analysis of the results of the Specification stage and lasted about 3 weeks. The Standardisation meeting took place with 11 judges and the project coordinator. All judges had attended the Familiarisation stage, but two of them did not participate in the Specification meeting as mentioned in section 4. However, they were given the corresponding interim reports and were asked to read them carefully, in order to be familiar with the results of that phase.

Special care was taken in the design of the material used during the meeting. Based on the methodology of the Manual, the Standardisation meeting comprised 4 parts. First, Familiarisation tasks were organised on the first day, the results of which were described above.

Training with samples calibrated to the CEFR is the next phase of the Standardisation phase and samples illustrating the CEFR levels had to be chosen. The Council of Europe DVD was used for spoken samples, from which one sample was initially played for familiarisation purposes and then the judges were asked to estimate the CEFR level of the performance of two learners on the DVD. For writing, two essays were taken from Tankó (2004). This book contains writing samples as a result of the Exam Reform Teacher Support Project in Hungary ([www.examsreform.hu](http://www.examsreform.hu)). These essays were preferred over the Cambridge samples in the Council's CD of illustrative samples because judgements could be affected if panellists knew the suite they were taken from. The source of the samples was only revealed at the end of the rating task. For Reading, three items of the Matriculation Examination in Finland were used, which were taken from the Council of Europe CD. From the same CD, two DIALANG items were chosen. Similarly, oral and written samples were used from Trinity in order for Benchmarking to take place. Oral samples were the ones that would later be given to exam centres for familiarisation with exam content; therefore, it would make sense to use these samples which were collected as examples of the Trinity suites. Written samples were selected by the project coordinator bearing in mind that performance closer to pass score needed to be preferred because of the standard setting methodology which concentrates on borderline performance. However, this was not always the case for the Portfolio samples as can be seen in Appendix 7 (p. 90).

Apart from samples, printed material had to be prepared for both Training and Benchmarking. First of all, rating forms like the one in Appendix 5 (p.88) were used for all skills on Form B3 of the Manual (Council of Europe, 2003:80). Along with the forms, the judges were given a booklet containing the descriptors from Tables 5.4, 5.5 and 5.8 of the Manual, as well as the page numbers of language activities scales in the CEFR (Council of Europe, 2001:Ch. 4) for all skills, Table 1 from the CEFR (Global

scale) and Tables 4.3, 4.4 and 4.5 from the Manual for aspects of language competence for reception, interaction and production.

For the final part of the phase, that is, Standard Setting, an appropriate method had to be chosen. The variant of the Angoff method suggested in the Manual (Council of Europe, 2003:91) was found convenient for this context, however, an investigation of the literature revealed that the Extended Angoff method (Hambleton & Plake, 1995) was also very suitable for the following reasons:

1. it deals with performance assessment, which is very relevant to Trinity exams
2. the task appeared to be quite straightforward
3. it did not involve use of many performance samples since this was already done during benchmarking and would make the task rather tedious
4. a lot of group discussion is suggested after the end of standard setting, which would allow for more in-depth understanding of the process
5. it suited the time availability, given that the Standardisation stage would need to include time for Training and Benchmarking

For this part of Standardisation, another rating form was created, which asked judges to decide on the cut-off scores in relation to the CEFR (see Appendix 6, p. 89), along with a booklet containing CEFR scales for language activities and competences as was the case with Training and Benchmarking. A questionnaire was prepared for the final day, in which judges were asked to report on the aspects of the project they liked or disliked. They were also asked to indicate their level of confidence in the standards they had set, as suggested in the literature (Hambleton, 2001:107; Hambleton & Plake, 1995).

#### **4.2.2 During the meeting**

The Familiarisation phase was conducted without any particular difficulty, since the same tasks and descriptors were used as in the previous phases of the project. The process was identical, with judges working on placing the descriptors at the CEFR levels, followed by a plenary discussion on how they approached the task and which descriptors were difficult to be assigned the correct level. During this meeting as well as the previous ones, a laptop and a data projector were used and correct answers were displayed on the screen, stimulating discussion about level misplacement. As stated previously, judges were also given feedback in the form of the ruler map generated by FACETS for the Global scale. This triggered group discussion regarding the group's scaling of the descriptors as it appeared on the FACETS output.

Training and Benchmarking followed more or less the same process. For oral performances, DVD files were broadcast on the screen and written samples were given in folders. Listening files from DIALANG were also played on a Media Player programme and powerful speakers were connected to the laptop. Judges were asked to use the material described above and individually assign CEFR levels to learners' performance. This should be accompanied by justification based on considering the CEFR scales. Since the Manual does not specify how many scales judges should look at in order to justify their level placement, the Trinity judges agreed to quote at least three different scales in their justification from those in their booklet. The issue of which scales should be consulted was discussed among the judges and agreed before the level placement. After judging the performance and assigning CEFR levels, group discussions followed, during which participants were asked to explain to the group how they approached the task and how they made decisions.

Finally during the Standard Setting phase, judges were requested to fill in the form in Appendix 6, where they indicated the CEFR level of imaginary candidates who receive each of the three scores awarded for each GESE Grade and ISE level. A plenary discussion followed during which the judges were asked to explain how they arrive at the cut-off scores in relation to the CEFR. Then a second round of judgments took place, where judges were asked to consider the borderline pass person for each GESE Grade and ISE level. Providing a second round of judgements is also adopted by Hambleton and Plake (1995:45) as well as asking judges to indicate their confidence level in the standards they have set; the Trinity judges indicated their confidence in a questionnaire at the end of the session as stated above. During decision making, judges could use the booklet with the CEFR scales described in the previous subsection.

### **4.2.3 After the meeting**

The outcome of the Standardisation meeting was a variety of quantitative and qualitative data which are discussed in detail in subsequent sessions. First of all, the Standardisation phase plenary discussions were digitally recorded in case I would like to clarify issues that had to do with decision making. Furthermore, this data would be the focus of my PhD research on how users of the Framework make decisions in the context of linking exams to the CEFR. Additional data were the written comments that judges were invited to include in the rating forms.

Apart from qualitative data, quantitative ones were collected. The ratings of the learners' performance during Training and Benchmarking, as well as the cut-off scores during Standardisation were converted into numbers and inserted into SPSS for statistical analysis of the reliability of judgements and in order to summarise decisions. The rationale for converting levels into numbers was the following: Initially, I examined all levels assigned by the judges in order to identify their range. Table 4.7 summarises those levels and their corresponding number. During the sessions, judges were not always confident that performance would fall within one CEFR band only. For example, there were cases where they would see elements of both the ordinary A1 descriptor and elements of the 'plus descriptor'. In this case I converted the use of two levels (A1 and A1+) into 3, one higher than performance that would only fall within A1 (which I converted into 2). When performances were either clearly at the 'plus' level or had elements of two consequent bands, then they were given the next number; for example A1/A2 and A1+ in the table below correspond to 4. It was decided that these two bands would be converted into the same number because there are CEFR scales without a 'plus' level, resulting in choice of A1/A2 as indication of a 'plus' performance. Finally, cases of performance which were clearly at the 'plus' level, but also exhibited characteristics of the next CEFR band were given an additional point, for example A1+/A2 was converted into 5.

It should be stressed that the final scale presented in Table 4.7 should only be considered as an arbitrarily constructed scale for measurement purposes of the present report and does not suggest any equal distance between the different levels and sublevels. Based on that scale, scores in the rating forms by judges were inserted in SPSS and descriptive statistics were calculated. More specifically, the reader will find measures of dispersion and central tendency (Alderson et al., 1995:92-95) reported in the following sections.

To conclude, the discussion of the results will focus on statistical analysis of judgements as well as the comments that the judges wrote on their rating forms, thus

giving a more complete picture of the Trinity Standardisation phase and the final decision making.

*Table 4.7 Conversion of CEFR levels into quantitative data*

<b>CEFR level</b>	<b>Scale</b>
Below A1	1
<b>A1</b>	<b>2</b>
A1/A1+	3
A1/A2	4
A1+	4
A1+/A2	5
<b>A2</b>	<b>6</b>
A2/A2+	7
A2/B1	8
A2+	8
A2+/B1	9
<b>B1</b>	<b>10</b>
B1/B1+	11
B1/B2	12
B1+	12
B1+/B2	13
<b>B2</b>	<b>14</b>
B2/B2+	15
B2/C1	16
B2+	16
B2+/C1	17
<b>C1</b>	<b>18</b>
C1/C1+	19
C1/C2	20
C1+	20
C1+/C2	21
<b>C2</b>	<b>22</b>

### 4.3 Training

Training samples judgements are analysed in this section by employing a number of indices and descriptive statistics. The aim of this section is to investigate reliability of judgements and make inferences on how judges rated already calibrated samples, using the CEFR scales as their marking criteria.

#### 4.3.1 Investigating consistency and agreement

Table 4.8 summarises coefficients indicating consistency and agreement. Spearman correlations along with Cronbach alpha are presented. The intraclass correlation coefficient, ICC, also reported in a CEFR-related study (Generalitat de Catalunya, 2006:62) is calculated in order to demonstrate how the average rater agreed with all others. Nichols (2006) offers brief guidelines on how to calculate ICC with SPSS, which were followed here. A more detailed discussion of ICC can be found in McGraw & Wong (1996). In Table 4.8 the ICC two-way mixed model was used and average measures for exact agreement are reported. Finally, Kendall's W has been used for investigating rater agreement in CEFR-related studies (Generalitat de

Catalunya, 2006:112; Kaftandjieva & Takala, 2002). As the SPSS Help Tool explains, Kendall's W can be interpreted as a coefficient of agreement among raters. Each case (row) is a rater and each variable (column) is an item being rated. The coefficient W ranges from 0 to 1, with 1 indicating complete inter-rater agreement, and 0 indicating complete disagreement among raters. All indices show high consistency and agreement among judges.

Table 4.8 Agreement and consistency of judges-training sessions

Stage	Inter-rater reliability			Alpha	ICC**	W**
	Mean*	Min	Max			
Training	0.883	0.684	0.996	0.983	.982	.870

\* Average using Fisher's Z-transformation

\*\* Statistically significant at level  $p \leq .01$

### 4.3.2 Investigating the rating process

Interesting conclusions can be drawn based on the way judges used the CEFR scales and rated calibrated samples, by looking at descriptive statistics. Summary statistics of judgements during Training are presented in Table 4.9 and will be discussed for all phases of the Standardisation stage similarly. The first column introduces the learners or items that the judges were asked to rate. In the next columns descriptive statistics of the judges' ratings are given. Measures of central tendency are calculated, such as the mean, that is, the arithmetic average of ratings, the median, which is the middle score and the mode which is the most frequent score. Measures of dispersion include the standard deviation, abbreviated to SD, which shows how much on average scores deviate from the mean and range which is the difference between the highest and the lowest rating. The minimum and the maximum ratings can also be found in the last two columns.

Table 4.9 Training results-summary statistics

Learner/Item	Mean	Median	Mode	SD	Range	Min	Max
Marcel (CoE DVD)	6.73	6	6	1.62	4	6	10
Doris (CoE DVD)	15.27	16	16	1.01	2	14	16
Essay Sample 1(Tanko, 2004:152)	9.91	10	10	1.03	5	8	13
Essay Sample 2 (Tanko, 2004:167)	13.73	14	14	0.47	1	13	14
Finnish reading item 1 (CoE CD)	18	18	18	2.68	8	14	22
Finnish reading item 2(CoE CD)	19.27	18	18	2.24	6	16	22
Finnish reading item 3(CoE CD)	16.91	14	14	3.62	8	14	22
DIALANG listening item 1(CoE CD)	16.73	17	18	1.42	4	14	18
DIALANG listening item 2(CoE CD)	7.64	6	6	2.38	7	6	13

The two first samples were two learners from the CoE DVD, namely Marcel and Doris who are described by North and Hughes (2003: 9,11) as 'a strong A2 (but not A2+)' and 'a strong performance' for B2 respectively. Apart from 2 judges, who preferred B1, 9 rated Marcel as A2 and the A2 descriptor from Table 5.4 of the Manual was quoted to justify decisions. It should be stressed that judges were asked to rate Marcel holistically whereas Doris was judged analytically; there was also one sample that was used as a warm up, without formal rating. Doris was also rated as a

strong performance, with the majority giving B2+ and referring to the B2+ descriptors from North and Hughes (2003: 26), which were included in the judges' handout.

The written samples aim at A2/B1 and B2 learners respectively (Tanko, 2004:21). The Trinity judges used Table 5.8 from the Manual (Council of Europe, 2003:82) to assign levels. For the second essay the range and standard deviation were very low, suggesting high agreement. B2 descriptors from Table 5.8 were found to describe performance in essay 2 accurately, even though some judges felt that not everything in the B2 descriptors was fulfilled. For essay 1 a wider range was observed, with elements from descriptors in A2 and B1 judged to be applicable to this sample.

For reading, Tasks 1, 2 and 3 from the Matriculation Examination in Finland in the CoE CD were used. Levels reported by the test providers are summarised in Table 4.10.

*Table 4.10 Items from the Matriculation Examination in Finland*

<b>CEFR Level</b>	<b>Task 1</b>	<b>Task 2</b>	<b>Task 3</b>
Text likely to be comprehensible by learner at CEFR level	B2	B2/C1	B2
Item level estimated	C1	B2	B2
Task level estimated	C1	B2	B2

The rationale given by the test constructors is based on the scale for Reading for Information and Argument (Council of Europe, 2001:70), from which the B2 and C1 descriptors are quoted for tasks 1 and 2 and B1 and B2 descriptors for B2. The Trinity judges rated the items somewhat higher, however the range of judgements was wide, especially for Tasks 1 and 3, which could be explained by the fact that according to their written comments, most of the judges were influenced by the C1 descriptor of the overall reading comprehension scale, which talks about lengthy and complex texts; despite the fact that the texts were not perceived as lengthy, the notion of complexity seemed to have affected judges. At the same time some judges reported difficulty with the use of the reading descriptors because as one judge put it, 'the descriptors poorly describe what learners can understand'. It should also be mentioned that because of the Trinity approach of integrating reading into writing, such an approach could have resulted in a variety of assigned levels.

The two DIALANG Listening items from the CoE CD were at B2 and A2 level. The Trinity judges appeared to be on par with that; they found both B2 and B2+ of the overall listening comprehension scale more relevant for item 1 and A2 and A2+ for item 2, even though some of the judges suggested C1 and one judge B1+/B2, after consulting the scale for Listening to Announcements and Instructions (Council of Europe, 2001:67).

### **4.3.3 Conclusion**

To conclude, Training with standardised samples appears to confirm a common understanding of the CEFR by the group; apart from some discrepancy in judgements for the reading items, the levels assigned by the Trinity group were in agreement with the test constructors' claims.

## **4.4 Benchmarking**

Benchmarking judgements are analysed in this section by employing an identical analysis to the previous section, since the Benchmarking task, as noted in the Manual,

is organised similarly to Training. This time however, local samples are chosen, that is, samples from Trinity candidates which were included in the DVD circulated to exam centres illustrating the content of GESE and ISE. These candidates were rated using the CEFR scales, but judges also rated performance using the Trinity marking criteria. Using the Trinity criteria had a dual purpose: first it would help identify borderline candidates, which is essential for the standard setting session and second, it was an opportunity for Trinity to check whether this panel of very experienced examiners would agree with the suggested marks by the test developers. The use of the marking criteria is not the primary focus of the present report, however all ratings are included in Appendix 7 (p. 90) because they will be mentioned later when ratings using the CEFR scales are analysed.

As with the Training section, I will discuss consistency and agreement of the panel and ratings of samples regarding the following: first, GESE in subsections 4.4.1-4.4.6, second, the speaking component of ISE in subsections 4.4.7-4.4.8 and finally, the writing component of ISE in subsections 4.4.9-4.4.11.

#### 4.4.1 Investigating consistency and agreement for the GESE suite

Table 4.11 summarises coefficients of consistency and agreement; details on these coefficients can be found in subsection 4.3.1. All judges watched DVD samples from Trinity Grades and provided ratings of the candidates' performance using the CEFR scales. Group discussion followed the rating of the Grades of each GESE stage as will be explained in detail in subsections 4.4.2-4.4.5. In general, as Table 4.11 shows, internal consistency of judgements and agreement were very high. Indices were calculated by converting judges' ratings using Table 4.7 and then using SPSS; this applies to all sets of ratings in the Benchmarking sessions. Note that Kendal's W had to be limited to the scale that was used by all raters, because it could not be calculated by SPSS when missing values occurred in the data. The reason why not all scales were used by the same amount of judges is explained in the following subsections.

*Table 4.11 Agreement and consistency of judges-GESE benchmarking*

Stage	Inter-rater reliability			Alpha	ICC**	W**
	Mean*	Min	Max			
Benchmarking	0.954	0.888	0.983	0.994	.932	.978***

\* Average using Fisher's Z-transformation

\*\* Statistically significant at level  $p \leq .01$

\*\*\*Only overall spoken interaction scale

#### 4.4.2 Investigating the rating process for GESE Initial Grades

At the beginning of this Benchmarking session, judges watched a DVD sample from Grade 1 which they did not rate. This was done as warm-up activity, in order to show judges how they were expected to work. It also initiated a lively discussion on which scales to use, since, as has already been stated, the Manual does not indicate how many scales should be consulted in order to provide adequate justification for CEFR level assignment to a performance. It was decided that 3 scales should be the minimum number, but as it will be shown later, the higher the Grade, the more scales were used.

Table 4.12 summarises judgements for Initial Grades. The name of the candidate can be seen in the first column. The second column refers to the CEFR scales used by

Trinity panellists in order to judge performance along with their page number in the CEFR and the third column shows the number of judges using each scale. Columns 4-8 contain the measures of dispersion and central tendency used in the previous subsection. The minimum and maximum ratings for each category are given in the last two columns.

Three DVD samples were broadcast and the judges rated them using the CEFR scales (there was one judge that could not attend the first sample). Table 4.12 shows an overall satisfactory agreement: The Grade 1 candidate was rated as A1, then the Grade 2 candidate appeared to exhibit characteristics from both A1 and A2, whereas the Grade 3 candidate showed a strong A2 performance.

In their written comments, the judges indicated among others that Carla “can interact in a simple way and respond to simple statements in areas of immediate need” (Overall Spoken Interaction Scale), “can follow short, simple directions and can understand questions and instruction addressed clearly and slowly” (Understand NS Interlocutor) and “can follow speech which is slow and carefully articulated” (Overall Listening Comprehension). Cristina was found to fulfil the A1 level description in the same scales and even demonstrates some of the characteristics of A2, such as “Can understand what is said clearly, slowly and directly to him/her in simple everyday conversation” (Understand NS Interlocutor) and “can give a simple description” (Overall Oral Production). Finally, Marco was seen as a clear A2 performance even with some elements of A2+ such as “Can interact with reasonable ease in structured situations and short conversations” (Overall Spoken Interaction).

*Table 4.12 Benchmarking Initial Grades-summary statistics*

<b>Learner</b>	<b>Scales</b>	<b>N</b>	<b>Mean</b>	<b>Median</b>	<b>Mode</b>	<b>SD</b>	<b>Range</b>	<b>Min</b>	<b>Max</b>
<b>Carla</b> <b>Grade 1</b>	Overall Sp. Inter. p.74	10	2	2	2	0	0	2	2
	Overall List. Comp. p.66	10	2	2	2	0	0	2	2
	Understand NS Interl. p.75	10	2	2	2	0	0	2	2
	Overall oral production p.58	1	2	2	2		0	2	2
<b>Cristina</b> <b>Grade 2</b>	Overall Sp. Inter. p.74	11	4	4	4	1.55	4	2	6
	Overall List. Comp. p.66	11	4.55	4	6	1.57	4	2	6
	Understand NS Interl. p.75	11	4.36	6	6	1.96	4	2	6
	Overall oral production p.58	10	4.2	5	6	1.99	4	2	6
<b>Marco</b> <b>Grade 3</b>	Overall Sp. Inter. p.74	11	7	7	6	1	2	6	8
	Overall List. Comp. p.66	11	6.73	6	6	1.35	4	6	10
	Understand NS Interl. p.75	11	7.18	8	8	0.98	2	6	8
	Overall oral production p.58	11	6.18	6	6	0.6	2	6	8

#### **4.4.3 Investigating the rating process for GESE Elementary Grades**

The Grade 4 candidate (see Table 4.13) demonstrated, according to the judges, elements of A2 and some B1 such as “Can reasonably fluently sustain a straightforward description” (Overall Oral Production). Some wider range of judgements for this candidate as opposed to others is observed, which according to judges’ comments, seemed to be because the candidate failed to understand a question by the examiner.

Monica was rated as a candidate demonstrating B1 elements, as well as some A2+. For example she was seen as being able to “enter unprepared into conversation on familiar topics” (Overall Spoken Interaction) and “Can understand the main points

of clear standard speech on familiar matters regularly encountered in work, school, leisure etc” (Overall Listening Comprehension).

Renate was evaluated as a good B1 performance, showing behaviour described both at B1 and B1+ such as “Can communicate with some confidence on familiar routine and non-routine matters related to his/her interests” (Overall Spoken Interaction). The judges also pointed out that the scale for Understanding a Native Speaker Interlocutor was hard to use, as the B2 descriptor mentions ‘noisy environment’ which is not supposed to be the case for an exam room.

*Table 4.13 Benchmarking Elementary Grades-summary statistics*

<b>Learner</b>	<b>Scales</b>	<b>N</b>	<b>Mean</b>	<b>Median</b>	<b>Mode</b>	<b>SD</b>	<b>Range</b>	<b>Min</b>	<b>Max</b>
<b>Susana Grade 4</b>	Overall Sp. Inter. p.74	10	6.8	6	6	1.4	4	6	10
	Overall List. Comp. p.66	8	8.25	8	8	1.16	4	6	10
	Understand NS Interlocutor p.75	11	7.91	8	8	1.45	4	6	10
	Sustained mon.: Descr. Exp. p.59	4	7.75	8	8	1.26	3	6	9
	Overall oral production p.58	11	8	8	8	1.26	4	6	10
<b>Monica Grade 5</b>	Overall Sp. Inter. p.74	11	9.91	10	10	0.7	2	9	11
	Overall List. Comp. p.66	7	10.1	10	10	0.38	1	10	11
	Understand NS Interlocutor p.75	11	10	10	10	0	0	10	10
	Sustained mon.: Descr. Exp. p.59	3	10	10	10	0	0	10	10
	Conversation p. 76	8	9.88	10	10	0.35	1	9	10
	Overall oral production p.58	10	9.8	10	10	0.63	2	8	10
<b>Renate Grade 6</b>	Overall Sp. Inter. p.74	11	11.5	11	11	0.93	3	10	13
	Overall List. Comp. p.66	9	10.9	11	10	0.93	2	10	12
	Understand NS Interlocutor p.75	9	11.3	12	10	1.41	4	10	14
	Sustained mon.: Descr. Exp. p.59	3	10	10	10	0	0	10	10
	Conversation p. 76	8	10.5	10	10	0.93	2	10	12
	Overall oral production p.58	9	11.3	12	10	1.41	4	10	14
	Analytical	1	12	12	12		0	12	12
	Global scale	1	14	14	14		0	14	14

#### **4.4.4 Investigating the rating process for GESE Intermediate Grades**

When rating candidates at the Intermediate stage, the judges decided to use more scales. This could have been the case either because the judges found the scales more relevant for more proficient candidates as they stated in the meeting, or because they were more familiar with the process and the scales after having used them for six samples. The judges also decided to use the qualitative aspects of spoken language use from Table 3 in the CEFR (Council of Europe, 2001:28-29) which is also included in the Manual as Table 5.5 (Council of Europe, 2003:79). These aspects correspond to the “Analytical” entries in Table 4.14, along with a holistic judgement based on these scales.

Table 4.14 Benchmarking Intermediate Grades-summary statistics

Learner	Scales	N	Mean	Median	Mode	SD	Range	Min	Max
<b>Michele Grade 7</b>	Overall Sp. Inter. p.74	10	13.8	14	14	0.63	2	12	14
	Overall List. Comp. p.66	6	13.3	14	14	1.63	4	10	14
	Understand NS Interlocutor p.75	9	14	14	14	0	0	14	14
	Sustained mon.: Descr. Exp. p.59	2	14	14	14	0	0	14	14
	Conversation p. 76	5	14	14	14	0	0	14	14
	Overall oral production p.58	6	14	14	14	0	0	14	14
	Analytical-overall	5	13.6	14	14	0.89	2	12	14
	Analytical-range	3	13.3	14	14	1.15	2	12	14
	Analytical-accuracy	4	13	13	12	1.15	2	12	14
	Analytical-fluency	5	14	14	14	0	0	14	14
	Analytical-interaction	4	13.5	14	14	1	2	12	14
	Analytical-coherence	4	14	14	14	0	0	14	14
	Global scale	1	14	14	14		0	14	14
<b>Paolo Grade 8</b>	Overall Sp. Inter. p.74	10	13.3	14	14	0.95	2	12	14
	Overall List. Comp. p.66	4	14	14	14	0	0	14	14
	Understand NS Interlocutor p.75	8	13.5	14	14	1.41	4	10	14
	Sustained mon.: Descr. Exp. p.59	2	10	10	10	0	0	10	10
	Conversation p. 76	5	12.4	14	14	2.19	4	10	14
	Overall oral production p.58	8	14	14	14	1.07	4	12	16
	Analytical-overall	7	13.3	14	14	1.5	4	10	14
	Analytical-range	5	14	14	14	0	0	14	14
	Analytical-accuracy	4	13.5	14	14	1	2	12	14
	Analytical-fluency	6	14	14	14	0	0	14	14
	Analytical-interaction	6	13.8	14	14	0.41	1	13	14
	Analytical-coherence	3	14	14	14	0	0	14	14
	Informal discussion p. 77	2	14.5	14.5	14	0.71	1	14	15
Information exchange p. 81	2	10	10	10	0	0	10	10	
Global scale	0								
<b>Cristian Grade 9</b>	Overall Sp. Inter. p.74	10	12.5	12	12	1.27	4	10	14
	Overall List. Comp. p.66	4	13.5	14	14	2.52	6	10	16
	Understand NS Interlocutor p.75	7	14	14	14	0	0	14	14
	Sustained mon.: Descr. Exp. p.59	1	14	14	14		0	14	14
	Conversation p. 76	6	12	12	10	2.19	4	10	14
	Overall oral production p.58	11	12.2	12	14	1.89	4	10	14
	Analytical-overall	4	12	12	12	1.63	4	10	14
	Analytical-range	5	11.4	12	10	1.34	3	10	13
	Analytical-accuracy	7	12	12	12	1.63	4	10	14
	Analytical-fluency	7	13	13	12	1	2	12	14
	Analytical-interaction	6	13.7	14	14	0.82	2	12	14
	Analytical-coherence	4	13	13	12		2	12	14
	Formal discussion p.78	1	14	14	14		0	14	14
Informal discussion p. 77	2	14	14	14		0	14	14	
Information exchange p. 81	0								
Global scale	1	12	12	12		0	12	12	

The Grade 7 candidate demonstrated, according to the judges, B2 behaviour, even though there were some opinions for B1+, for example accuracy. Michele could “understand in detail what is said to her in the standard spoken language” (Understanding a Native Speaker Interlocutor) even though, as mentioned above, the environment cannot be noisy as the descriptor states. It was also found, among others,

that the B2 Overall Spoken Interaction fits very well, even though participants were not sure about the “imposing strain” part of the descriptor and how they should interpret “personal significance of events” (“Can interact with a degree of fluency and spontaneity that makes regular interaction, and sustained relationships with native speakers quite possible without imposing strain on either party. Can highlight the personal significance of events and experiences, account for and sustain views clearly by providing relevant explanations and arguments.” (Council of Europe, 2001:74)).

Some wider discrepancy of judgements was observed for the Grade 8 candidate. Despite taking a higher Grade, Paolo was sometimes either at the same CEFR level or slightly lower than Michele. This might be because, as a judge noted in the written comments, he did not seem to initiate the discussion, and that might have resulted in a somewhat lower rating. However, the same judge noted that this lack of initiation was not the case for the Interactive task of the Interview, which is intended for the candidate to control the discussion.

The Grade 9 candidate caused a lot of discussion among the participants and never did he appear to be placed higher than the candidates at Grades 7 and 8. Many judges saw elements of a B1 or B1+ performance on the Overall Spoken Interaction scale, range and accuracy. The group attributed that to the candidate’s age as well as the high entrance in the GESE suite, that is, he should have sat a lower Grade. By looking back to the ratings using the Trinity criteria (p. 90), the majority of judges failed the candidate, even though the initial Trinity suggestion was a pass. It should be stressed here that candidates at the intermediate stage are assessed on three tasks and a compensatory approach is applied for the final mark. C is the pass mark.

Placing Christian at a lower CEFR level and also receiving an initial suggestion for an overall pass raised the issue of marking reliability or reconsideration of the level of the Intermediate Grades. In other words, if on the one hand the candidate should have passed (initially Trinity and four judges suggested an overall pass) then claiming that Grade 9 is B2 was an overestimation; on the other hand, if he should have failed, then that for some people he was a bare pass should be reconsidered, along with the reliability of marks awarded. This issue was agreed to be discussed further during the Standard Setting session.

#### **4.4.5 Investigating the rating process for GESE Advanced Grades**

Due to time constraints, the group decided that a sample from Grade 11 would be omitted from the Advanced stage session, since this could be replaced by a later sample from ISE III Interview, which has exactly the same structure.

The judgements and discussion on Gabriel’s performance (Table 4.15) revealed some interesting findings on the CEFR scales. The judges found Gabriel to be at the B2+/C1 range, but felt that there was an issue of maturity. In fact Gabriel was judged to be very proficient but his inability to provide a correct answer to the listening task<sup>1</sup> was attributed to his age. Written comments pointed out that the CEFR scales at C1 are not suitable for such a young learner, despite his advanced linguistic competence. This is a very interesting finding, because it indicates that apart from suggestions in the literature that the CEFR is not suitable for young learners at the lower levels (Carlsen & Moe, 2005), this seems to apply to highly proficient young learners, too

---

<sup>1</sup> The Advanced Grades contain a discrete listening task, during which the examiner reads 3 short texts and the candidate is expected to suggest possible endings for the first two or identify participants, contexts or settings for the third one (Trinity College London, 2005a: 43).

and it appears to confirm Hasselgreen's (2005:352) claim that there seems to be a C1 ceiling for young learners.

Bobi's performance was unanimously judged as a clear C2 performance and as can be seen in Table 4.15, he was given a distinction by all raters. One of the comments of the judges regarding the CEFR descriptors was that the scale for Addressing Audiences mentions "Can handle difficult and even hostile questioning" which cannot be the case in an exam situation.

Table 4.15 Benchmarking Advanced Grades-summary statistics

Learner	Scales	N	Mean	Median	Mode	SD	Range	Min	Max
<b>Gabriel Grade 10</b>	Overall Sp. Inter. p.74	10	17.4	18	18	0.84	2	16	18
	Overall List. Comp. p.66	9	16.4	16	18	1.67	4	14	18
	Understand NS Interlocutor p.75	2	17	17	16	1.41	2	16	18
	Sustained mon.: Descr. Exp. p.59	3	16	16	14	2	4	14	18
	Conversation p. 76	2	15	15	12	4.24	6	12	18
	Overall oral production p.58	9	16.4	16	18	1.67	4	14	18
	Analytical-overall	7	16.6	17	16	1.4	4	14	18
	Analytical-range	4	16	16.5	17	1.41	3	14	17
	Analytical-accuracy	3	15.7	16	14	1.53	3	14	17
	Analytical-fluency	4	16.3	16.5	14	1.71	4	14	18
	Analytical-interaction	5	17.2	18	18	1.1	2	16	18
	Analytical-coherence	2	16.5	16.5	16	0.71	1	16	17
	Formal discussion p.78	1	18	18	18		0	18	18
	Informal discussion p. 77	0							
	Information exchange p. 81	0							
	Addressing audiences p. 60	9	16.3	16	16	1.58	4	14	18
Global scale	1	14	14	14		0	14	14	
<b>Bobi Grade 12</b>	Overall Sp. Inter. p.74	10	22	22	22	0	0	22	22
	Overall List. Comp. p.66	11	22	22	22	0	0	22	22
	Understand NS Interlocutor p.75	3	22	22	22	0	0	22	22
	Sustained mon.: Descr. Exp. p.59	4	22	22	22	0	0	22	22
	Conversation p. 76	2	22	22	22	0	0	22	22
	Overall oral production p.58	8	22	22	22	0	0	22	22
	Analytical-overall	6	22	22	22	0	0	22	22
	Analytical-range	5	22	22	22	0	0	22	22
	Analytical-accuracy	5	22	22	22	0	0	22	22
	Analytical-fluency	5	22	22	22	0	0	22	22
	Analytical-interaction	5	22	22	22	0	0	22	22
	Analytical-coherence	5	22	22	22	0	0	22	22
	Formal discussion p.78	1	22	22	22		0	22	22
	Informal discussion p. 77	0							
	Information exchange p. 81	0							
	Addressing audiences p. 60	8	21.8	22	22	0.71	2	20	22
Global scale	1	22	22	22		0	22	22	

#### 4.4.6 Investigating consistency and agreement for ISE Interview

Table 4.16 shows the inter-rater reliability and internal consistency of the group which suggest high agreement among the group members. Details about the way indices were calculated can be found in subsection 4.4.1.

Table 4.16 Agreement and consistency of judges -ISE Interview benchmarking

Stage	Inter-rater reliability			Alpha	ICC**	W**
	Mean*	Min	Max			
Benchmarking	0.940	0.725	0.991	0.999	0.998	0.963***

\* Average using Fisher's Z-transformation

\*\* Statistically significant at level  $p \leq .01$

\*\*\*Only overall spoken interaction and overall listening comprehension scales

#### 4.4.7 Investigating the rating process for ISE Interview

Table 4.17 summarises judgements about Monika's performance, a candidate interviewed for ISE 0.

Table 4.17 Benchmarking ISE 0 Interview-summary statistics

Learner	Scales	N	Mean	Median	Mode	SD	Range	Min	Max
Monika ISE0	Overall Sp. Inter. p.74	11	8.27	8	8	0.63	2	7	9
	Overall List. Comp. p.66	8	9.75	10	10	0.38	1	9	10
	Understand NS Interlocutor p.75	11	9.09	9	10	1.05	3	7	10
	Sustained mon.: Descr. Exp. p.59	4	7.5	8	8	1	2	6	8
	Conversation p. 76	4	8	8.5	9	1.41	3	6	9
	Overall oral production p.58	10	8	8	8	1.73	4	6	10
	Analytical-overall	9	9.22	9	9	0.64	2	8	10
	Analytical-range	3	10	10	10	0	0	10	10
	Analytical-accuracy	2	9	9	8	1.41	2	8	10
	Analytical-fluency	3	9.67	10	10	0.58	1	9	10
	Analytical-interaction	2	9	9	8	1.41	2	8	10
	Analytical-coherence	3	9	9	8	1	2	8	10
	Formal discussion p.78	1	8	8	8		0	8	8
	Informal discussion p. 77	0							
	Information exchange p. 81	1	8	8	8		0	8	8
	Addressing audiences p. 60	1	9	9	9		0	9	9
	Monitoring and repair p. 65	1	12	12	12		0	12	12
Global scale	1	9	9	9		0	9	9	

Overall, her performance demonstrated, according to the group, mainly elements of A2+ descriptors as well as B1, with some fewer cases of A2. One of the judges noted in the written comments that Monica is very communicative and fluent, but still makes grammatical mistakes, with listening skills and vocabulary being her strongest areas. Another judge wrote that she is a higher than average ISE 0. Based on the wording of the CEFR scales, Monica was able to describe routines, past activities, personal experiences as mentioned in A2+ level (Sustained Monologue: Describing Experience), whereas the help of the interlocutor described in the Overall Spoken Interaction A2+ descriptor (Council of Europe, 2001: 74) was not seen as necessary for her (Can interact with reasonable ease in structured situations and short conversations, provided the other person helps if necessary).

The majority of judgements about the ISE I candidate in Table 4.18 were in the B1 area, with some B1+ placements and two occasions of A2+/B1. Virginie's oral production was described, according to the panel, very accurately by the B1 descriptor

“Can reasonably fluently sustain a straightforward description of one of a variety of subjects within his/her field of interest, presenting it as a linear sequence of points”(Council of Europe, 2001: 58). The B1 descriptor for Overall Spoken Interaction (Council of Europe, 2001: 74) was also found relevant to describe her performance, even though there is no evidence of the context of travelling in her performance. The B1 descriptor from the scale for Understanding a Native Speaker Interlocutor was also chosen, even though Virginie does not ask for repetition.

*Table 4.18 Benchmarking ISE I Interview-summary statistics*

<b>Learner</b>	<b>Scales</b>	<b>N</b>	<b>Mean</b>	<b>Median</b>	<b>Mode</b>	<b>SD</b>	<b>Range</b>	<b>Min</b>	<b>Max</b>
<b>Virginie ISE I</b>	Overall Sp. Inter. p.74	10	10.1	10	10	0.33	1	10	11
	Overall List. Comp. p.66	8	11	11	10	1.07	2	10	12
	Understand NS Interlocutor p.75	10	10	10	10	0	0	10	10
	Sustained mon.: Descr. Exp. p.59	0							
	Conversation p. 76	3	9.667	10	10	0.58	1	9	10
	Overall oral production p.58	9	9.778	10	10	0.71	2	8	10
	Analytical-overall	10	10.3	10	10	0.67	2	10	12
	Analytical-range	5	10.4	10	10	0.89	2	10	12
	Analytical-accuracy	3	10.67	10	10	1.15	2	10	12
	Analytical-fluency	4	10.5	10	10	1	2	10	12
	Analytical-interaction	4	10.5	10	10	1	2	10	12
	Analytical-coherence	4	10.5	10	10	1	2	10	12
	Formal discussion p.78	1	10	10	10		0	10	10
	Informal discussion p. 77	0							
	Information exchange p. 81	0	10	10	10		0	10	10
	Addressing audiences p. 60	1	10	10	10		0	10	10
Global scale	2	11	11	10	1.41	2	10	12	

The majority of judgements about Mathilde’s performance were at B2 level with some at B1+ and one occurrence of B1 (Table 4.19). It appeared that her fluency, which as Table 4.19 points out was rated slightly lower, affected judgements for some lower level placements. The judges felt the Overall Spoken Interaction scale described her performance accurately (Council of Europe, 2001: 74). Some judges also expressed their concern that the B2 descriptor for Overall Oral Production (Council of Europe, 2001: 58) does not seem to be very relevant in this exam context, even though 5 judges included it in their justification of their decision.

Atahan, as can be seen in Table 4.20 was judged as C1 learner, with some B2+ and two B2 placements, one however was because the corresponding scale (Information Exchange) does not provide a description for C1 and C2 (Council of Europe, 2001: 81). One judge also noted that the candidate’s poor pronunciation might have affected her impression of his overall level. It should be stressed that one of the judges who had examined Atahan stressed that his pronunciation was not really a problem when interacting with him, which might support C1 decisions over B2+. Despite some language errors, the judges found that his range (analytical criteria) was broad and that C1 descriptors for Overall Oral Production and Overall Spoken Interaction applied.

Table 4.19 Benchmarking ISE II Interview-summary statistics

Learner	Scales	N	Mean	Median	Mode	SD	Range	Min	Max
<b>Mathilde ISE II</b>	Overall Sp. Inter. p.74	10	13.9	14	14	0.33	1	13	14
	Overall List. Comp. p.66	6	13.5	14	14	0.89	2	12	14
	Understand NS Interlocutor p.75	7	14	14	14	0	0	14	14
	Sustained mon.: Descr. Exp. p.59	1	13	13	13		0	13	13
	Conversation p. 76	4	13.75	14	14	1.26	3	12	15
	Overall oral production p.58	5	14	14	14	0	0	14	14
	Analytical-overall	8	13.88	14	14	0.35	1	13	14
	Analytical-range	5	13.2	14	14	1.1	2	12	14
	Analytical-accuracy	4	13	13	12	1.15	2	12	14
	Analytical-fluency	5	12.8	14	14	1.79	4	10	14
	Analytical-interaction	5	13.2	14	14	1.1	2	12	14
	Analytical-coherence	5	13.6	14	14	0.89	2	12	14
	Formal discussion p.78	2	14	14	14	0	0	14	14
	Informal discussion p. 77	0							
	Information exchange p. 81	1	12	12	12		0	12	12
	Addressing audiences p. 60	1	14	14	14		0	14	14
Global scale	2	14	14	14	0	0	14	14	

Table 4.20 Benchmarking ISE III Interview-summary statistics

Learner	Scales	N	Mean	Median	Mode	SD	Range	Min	Max
<b>Atahan ISE III</b>	Overall Sp. Inter. p.74	11	17.73	18	18	0.48	1	17	18
	Overall List. Comp. p.66	8	18.38	18	18	2.82	8	14	22
	Understand NS Interlocutor p.75	7	18	18	18	0	0	18	18
	Sustain. mon.: Descr. Exp. p.59	4	18	18	18	0	0	18	18
	Conversation p. 76	1	18	18	18		0	18	18
	Overall oral production p.58	7	18.14	18	18	0.98	3	17	20
	Analytical-overall	6	18	18	18	0	0	18	18
	Analytical-range	6	17.67	18	18	0.82	2	16	18
	Analytical-accuracy	6	16.33	16	16	1.51	4	14	18
	Analytical-fluency	5	17.6	18	18	0.89	2	16	18
	Analytical-interaction	6	18	18	18	0	0	18	18
	Analytical-coherence	5	17.6	18	18	0.89	2	16	18
	Formal discussion p.78	2	18	18	18	0	0	18	18
	Informal discussion p. 77	1	16	16	16		0	16	16
	Information exchange p. 81	1	14	14	14		0	14	14
	Identifying cues p.72	2	18	18	18	0	0	18	18
	Monitoring and repair p. 65	1	18	18	18		0	18	18
	Addressing audiences p. 60	3	17	17	16	0.71	2	16	18
	Global scale	2	18	18	18	0	0	18	18

#### 4.4.8 Investigating consistency and agreement for ISE I and II Written

Due to time constraints, the panel was divided into two groups of five and six participants for judging written performance. ISE Levels I and II were rated by six judges who were asked to assign CEFR levels to written samples (ISE 0 and ISE III are discussed in the following section). In particular, the samples were a Controlled Written Exam paper containing 2 tasks (Trinity College London, 2005b: 10) and a

Portfolio containing 3 tasks (Trinity College London, 2005b: 8) for each of the two levels.

During the calculation of group agreement and consistency it was realised that because two raters used a limited number of scales in the rating forms, correlations for these two were not in their majority statistically significant; for this reason these raters, Claudia and Lora are excluded from the results in Table 4.21. However, it was decided that that their judgements would be included in the ratings in subsection 4.4.9 below, provided that they were within one CEFR level, or 4 points on the conversion scale of Table 4.7, because the rest of the group were very consistent as shown in Table 4.21.

*Table 4.21 Agreement and consistency of judges-ISE I and II Written benchmarking*

Stage	Inter-rater reliability			Alpha	ICC**	W**
	Mean*	Min	Max			
Benchmarking	0.913	0.780	1	0.933	0.922	0.922***

\* Average using Fisher's Z-transformation

\*\* Statistically significant at level  $p \leq .01$

\*\*\*Only scales used by all judges

#### **4.4.9 Investigating the rating process for ISE I and II Written**

Table 4.22 summarises judgements about ISE I. As can be seen in the N column, judges tended to use the analytic criteria which were taken from Table 5.8 in the Manual (Council of Europe, 2003: 82). The overall category in these criteria has elements from the Overall Written Production and Overall Written Interaction scales (Council of Europe, 2001: 61, 83), which probably explains the infrequent use of these scales in some cases.

Because the exam specifications state that Reading is integrated into the Controlled Written exam, in particular Task 1, the judges decided to investigate the Reading scales as well. They found some description of A2+ and B1 from the Overall Reading Comprehension scale (Council of Europe, 2001: 69) to be appropriate for Maria, as well as the Processing Texts scale (Council of Europe, 2001: 96). The judges noted that they did not find the Reading scales very relevant and this could be because of the integrated nature of reading; as one judge wrote, the candidate does not show full evidence of understanding a lot of the text. However, it is not clear in this comment whether this is the case because of the candidate's poor performance or the nature of the test. There were also comments on the candidate's failure to fulfil the task in Question 1, however, this candidate was given a C, that is a pass. This might also have implications for marking reliability. Judgements on Maria's writing ability ranged from A2 to B1, based on the analytic marking criteria from Table 5.8 in the Manual (Council of Europe, 2003: 82). Very high agreement was also observed regarding Flavia's performance in the Portfolio, where almost all judgements rated her performance at B1, with only one A2 and one A2+/B1 placement.

Table 4.22 Benchmarking ISE I Written-summary statistics

Learner	Scales	N	Mean	Median	Mode	SD	Range	Min	Max
<b>Maria ISE I Controlled Written</b>	Overall Reading p. 69	4	8.75	8.5	8	0.96	2	8	10
	Reading corresp. p.69	0							
	Reading for orientat. p.70	1	6	6	6		0	6	6
	Reading for inf. & arg. p.70	0							
	Processing text p.96	2	8.5	8.5	8	0.71	1	8	9
	Overall Written Prod. p.61	0							
	Overall Written Inter. p.83	1	10	10	10		0	10	10
	Correspondence p. 83	0							
	Global Writ. p.82 Manual	3	10.7	10	10	1.15	2	10	12
	Analytical-overall	2	9.5	9.5	9	0.71	1	9	10
	Analytical-range	3	9.33	10	10	1.15	2	8	10
	Analytical-coherence	4	9	10	10	2	4	6	10
	Analytical-accuracy	4	8.5	9	10	1.91	4	6	10
	Analytical-description	3	10	10	10	0	0	10	10
	Analytical-argument	3	8.67	10	10	2.31	4	6	10
	Strategic comp. Table 4.3	3	6.67	6	6	1.15	2	6	8
Reports and essays p.62	1	12	12	12		0	12	12	
<b>Flavia ISE I Portfolio</b>	Overall Reading p. 69	1	10	10	10		0	10	10
	Reading corresp.p.69	2	10	10	10	0	0	10	10
	Reading for orient. p.70	0							
	Reading for inf. & arg. p.70	0							
	Processing text p.96	0							
	Overall Written Prod. p.61	0							
	Creative writing p.62	2	10	10	10	0	0	10	10
	Overall Written Inter. p.83	4	10	10	10	0	0	10	10
	Correspondence p. 83	4	9.75	10	10	0.5	1	9	10
	Global Writ. p.82 Manual	4	10	10	10	0	0	10	10
	Analytical-overall	4	10	10	10	0	0	10	10
	Analytical-range	4	10	10	10	0	0	10	10
	Analytical-coherence	4	9	10	10	2	4	6	10
	Analytical-accuracy	4	10	10	10	0	0	10	10
	Analytical-description	4	10	10	10	0	0	10	10
	Analytical-argument	3	10	10	10	0	0	10	10
Notes, mess. & forms p.84	1	10	10	10		0	10	10	

Judging Andrea's reading ability was not easy for judges, as they found only some applicability of the B1 Overall Reading Comprehension. One judge pointed out that the candidate misunderstood the task and did not summarise the writer's views, but gave his own views, therefore the task was not achieved. Another judge wrote that evidence of reading ability is practically non-existent and could not place him in a scale. Comments also stressed that B2 descriptors do not describe the type of reading task demonstrated here as there is no real evidence that the candidate has read the article with understanding. These comments suggest that the CEFR scales are not very suitable for assessing reading in an integrated way. In addition, the score obtained by the candidate (B) in combination with comments that the task was not fulfilled could possibly raise again an issue of marking reliability.

Andrea's written performance in Table 4.23 caused a wide range of judgements which appears to have been the result of judges using different CEFR scales as was also pointed out subsection 4.4.8 in relation to the calculation of judges' agreement

and consistency. For example, a judge who looked at the Overall Written Interaction scale (Council of Europe, 2001: 83) and the Reports and Essays scale (Council of Europe, 2001: 62) estimated Andrea's performance at B2, whereas a judge using the analytical criteria found evidence of A2 and B1. This might have implications for the writing scales, which have also been reported as causing discrepancies in judgements (Kaftandjieva & Takala, 2002).

Table 4.23 Benchmarking ISE II Written-summary statistics

Learner	Scales	N	Mean	Median	Mode	SD	Range	Min	Max
<b>Andrea ISE II Controlled Written</b>	Overall Reading p. 69	4	10.5	10	10	1	2	10	12
	Reading corresp. p.69	0							
	Reading for orient. p.70	0							
	Read. for inf. & arg. p.70	2	12	12	10	2.83	4	10	14
	Processing text p.96	1	14	14	14		0	14	14
	Overall Written Prod. p.61	0							
	Overall Written Inter. p.83	4	14	14	14	0	0	14	14
	Correspondence p. 83	0							
	Global Writ. p.82 Manual	2	10	10	10	0	0	10	10
	Analytical-overall	2	11.5	11.5	9	3.54	5	9	14
	Analytical-range	2	11	11	10	1.41	2	10	12
	Analytical-coherence	2	12	12	10	2.83	4	10	14
	Analytical-accuracy	2	9	9	6	4.24	6	6	12
	Analytical-description	1	6	6	6		0	6	6
Analytical-argument	1	14	14	14		0	14	14	
Strategic comp. Table 4.3	0								
Reports and essays p.62	3	14	14	14	0	0	14	14	
<b>Alberto ISE II Portfolio</b>	Overall Reading p. 69	0							
	Reading correspond. p.69	0							
	Reading for orient. p.70	0							
	Read. for inf & arg. p.70	0							
	Processing text p.96	1	14	14	14		0	14	14
	Overall Writ. Prod. p.61	0							
	Creative writing p.62	1	14	14	14		0	14	14
	Overall Written Inter.p.83	2	12	12	10	2.83	4	10	14
	Correspondence p. 83	3	12.7	12	12	1.15	2	12	14
	Global Writ. p.82 Manual	4	13	14	14	2	4	10	14
	Analytical-overall	4	15	14	14	2	4	14	18
	Analytical-range	4	14	14	14	0	0	14	14
	Analytical-coherence	5	14.8	14	14	1.79	4	14	18
	Analytical-accuracy	5	14.8	14	14	1.79	4	14	18
Analytical-description	5	14.8	14	14	1.79	4	14	18	
Analytical-argument	3	14	14	14	0	0	14	14	
Notes, mess., forms p.84	0								

Discrepancies are also observed regarding Alberto's Portfolio, with judgements from B1 to C1. A judge who looked at the analytical criteria suggested C1 but noticed that the B2 and C1 descriptors refer more to academic and work orientated writing. Perhaps this might suggest that ISE II Portfolio tasks could include such tasks in order to elicit such performance. Finally one judge pointed out that the third task appears to be plagiarised, which has very important implications for the use of Portfolio

assessment in a high-stakes context. Portfolio assessment has been reported to have positive impact in the classroom (Gottlieb, 1995; Klenowski, 2000; Tsagari, 2005), but it might be problematic in large scale assessment (Alderson & Banerjee, 2001; Callahan, 1995; Hamp-Lyons & Condon, 1993, 1999; Newman & Smolen, 1991), especially since in the Trinity context it counts towards 20% of the final mark. It should be noted that the candidate received a B for this part of the Portfolio.

#### 4.4.10 Investigating consistency and agreement for ISE 0 and III Written

ISE Levels 0 and III were rated by the remaining five judges who were asked to assign CEFR levels to written samples. Again, the samples were a Controlled Written Exam paper containing 2 tasks (Trinity College London, 2005b: 10) and a Portfolio containing 3 tasks (Trinity College London, 2005b: 8) for each of the two levels.

Similarly to previous sections, Table 4.24 summarises indices regarding judges' consistency and agreement, which are all at very high levels. Scales that were used by four or five judges were the ones included in the analysis. W coefficient was calculated by only including scales used by all five judges. The N column of Table 4.25 and Table 4.26 shows the number of judges using each scale.

*Table 4.24 Agreement and consistency of judges-ISE 0 and III Written benchmarking sessions*

Stage	Inter-rater reliability			Alpha	ICC**	W**
	Mean*	Min	Max			
Benchmarking	0.928	0.811	1	0.999	0.970	0.865***

\* Average using Fisher's Z-transformation

\*\* Statistically significant at level  $p \leq 0.01$

\*\*\*Only scales used by all judges

#### 4.4.11 Investigating the rating process for ISE 0 and III Written

Judgements for ISE 0 and ISE III are summarised in Table 4.25 and Table 4.26 respectively. Giorgio's reading ability was generally thought to fall within the A1/A2 range on the Overall Reading Comprehension scale (Council of Europe, 2001: 69), whereas A2 was more appropriate for Reading for Orientation (ibid: 70) and A2+ for Processing Text (ibid: 96). The majority of judgements for Giorgio's writing ability using the analytical written criteria (Council of Europe, 2003: 82) fell within the A2 band. One judge that found A1 relevant expressed the view that this was a weak candidate.

Judgements about Luca's performance in the ISE 0 Portfolio revealed high agreement; use of the analytical written criteria placed him at the A2 CEFR band.

Table 4.25 Benchmarking ISE 0 Written-summary statistics

Learner	Scales	N	Mean	Median	Mode	SD	Range	Min	Max
<b>Giorgio ISE 0 Controlled Written</b>	Overall Reading p. 69	5	4.6	5	6	1.67	4	2	6
	Reading correspondence p.69	0							
	Reading for orientation p.70	2	6	6	6	0	0	6	6
	Reading for info & argum. p.70	1	6	6	6		0	6	6
	Processing text p.96	4	7.5	8	8	1	2	6	8
	Overall Written Production p.61	0							
	Overall Written Interaction p.83	1	10	10	10		0	10	10
	Correspondence p. 83	1	6	6	6		0	6	6
	Global Written p.82 Manual	4	5.5	6	6	1	2	4	6
	Analytical-overall	0							
	Analytical-range	3	5.33	6	6	1.15	2	4	6
	Analytical-coherence	4	5	6	6	2	4	2	6
	Analytical-accuracy	4	5.5	6	6	1	2	4	6
	Analytical-description	1	4	4	4		0	4	4
	Analytical-argument	1	6	6	6		0	6	6
<b>Luca ISE 0 Portfolio</b>	Overall Reading p. 69	1	6	6	6		0	6	6
	Reading correspondence p.69	2	5	5	4	1.41	2	4	6
	Reading for orientation p.70	0							
	Reading for info & argum. p.70	0							
	Processing text p.96	0							
	Overall Written Production p.61	1	6	6	6		0	6	6
	Creative writing p.62	1	4	4	4		0	4	4
	Overall Written Interaction p.83	0							
	Correspondence p. 83	2	6	6	6	0	0	6	6
	Global Written p.82 Manual	5	6	6	6	0	0	6	6
	Analytical-overall	0							
	Analytical-range	4	6	6	6	0	0	6	6
	Analytical-coherence	4	6	6	6	0	0	6	6
	Analytical-accuracy	4	6	6	6	0	0	6	6
	Analytical-description	1	2	2	2		0	2	2
Analytical-argument	0								

Bernardo's reading ability offered an interesting finding: the Reading for Orientation scale resulted in lower level assignment than the Overall Reading Comprehension scale, as well as the Reading for Information and Argument and the Processing Text scales. This could suggest some problematic aspect of this scale or that some scales are more appropriate for some tasks than others. It could also be the case because the C1 and C2 levels do not have a description and are presented with the note "as B2"; it is not clear however how it can be possible to have the same description for these levels and whether then the B2 band is too ambitious for such a level. The candidate's written ability was estimated between the B2 and C1 levels, using the analytical criteria. Similarly, Stefania's performance fell between B2 and C1. One of the judges, who opted for B2, wrote that she found Stefania to be a weak candidate; therefore she opted for B2 and B2+ descriptors.

Table 4.26 Benchmarking ISE III Written-summary statistics

Learner	Scales	N	Mean	Median	Mode	SD	Range	Min	Max
<b>Bernardo ISE III Controlled Written</b>	Overall Reading p. 69	2	18	18	18	0	0	18	18
	Reading correspondence p.69	1	14	14	14		0	14	14
	Reading for orientation p.70	2	12	12	12	0	0	12	12
	Reading for info & argum. p.70	4	17	18	18	2	4	14	18
	Processing text p.96	3	18	18	18	0	0	18	18
	Overall Written Production p.61	0							
	Overall Written Interaction p.83	1	18	18	18		0	18	18
	Correspondence p. 83	3	13.33	14	10	3.06	6	10	16
	Global Written p.82 Manual	3	18	18	18	0	0	18	18
	Analytical-overall	1	18	18	18		0	18	18
	Analytical-range	5	16.4	18	18	2.19	4	14	18
	Analytical-coherence	5	16.4	18	18	2.19	4	14	18
	Analytical-accuracy	5	16.4	18	18	2.19	4	14	18
	Analytical-description	3	18	18	18	0	0	18	18
	Analytical-argument	5	15.6	14	14	2.19	4	14	18
Strategic comp. Table 4.3	0								
Reports and essays p.62	0								
<b>Stefania ISE III Portfolio</b>	Overall Reading p. 69	1	16	16	16		0	16	16
	Reading correspondence p.69	0							
	Reading for orientation p.70	0							
	Reading for info & argum. p.70	0							
	Processing text p.96	0							
	Overall Written Production p.61	0							
	Creative writing p.62	1	14	14	14		0	14	14
	Overall Written Interaction p.83	1	14	14	14		0	14	14
	Correspondence p. 83	2	14	14	14	0	0	14	14
	Global Written p.82 Manual	4	16	16	14	2.31	4	14	18
	Analytical-overall	0							
	Analytical-range	5	16	16	14	2	4	14	18
	Analytical-coherence	4	17	18	18	2	4	14	18
	Analytical-accuracy	4	14.5	14	14	1	2	14	16
	Analytical-description	5	15.6	14	14	2.19	4	14	18
Reports and essays p.62	1	16	16	16		0	16	16	
Analytical-argument	2	15	15	14	1.41	2	14	16	
Notes, messages and forms p.84	0								

#### 4.4.12 Conclusion

Benchmarking was an essential part of Standardisation because of the choice of the standard setting method selected for the session following benchmarking: since Angoff techniques involve estimating the level of the imaginary borderline person, the judges had the chance to work on Trinity samples and familiarise themselves with actual performance by candidates before deciding on the cut-off scores in relation to the CEFR which was the aim of the next session, i.e. Standard Setting. Even with the writing samples, the group discussion aimed to familiarise judges with all levels, even though each judge rated only samples from two ISE levels. Therefore, even though judges did not rate all samples, each of the two smaller teams was asked to present in detail to the other team the rationale for rating the corresponding samples.

From a research point of view, Benchmarking revealed a positive attitude by the panel. Judging the performance of learners was something that these panellists found very close to what they usually do as examiners and was commented on positively. This was a very important aspect of the Standardisation phase, because, as Hambleton (2001) suggests, in standard setting panels, judges should be asked to act similarly to their everyday practice. The judges also liked the combination of individual and group work during this phase, since they were asked to judge the performance of candidates using the CEFR scales and then discuss personal decision making with the group.

Benchmarking also revealed that judges might have preferences for different scales, which might hinder measuring agreement; for this reason, it might be more helpful for reporting results to reach consensus on the scales to be used before starting rating performance.

Finally, benchmarking local samples on the CEFR is an opportunity for the exam provider to collect expert judgements on the performance of candidates and the exam overall, not just in relation to the Framework. For example, Benchmarking here included rating performance using the Trinity marking criteria, discussing the portfolio assessment, etc. Therefore, Benchmarking can be used as another validation check of exam scores during the process of relating the exam to the CEFR.

## **4.5 Standard-setting**

The final stage of Standardisation, Standard-setting, is discussed in this section. Methodology is described in subsection 4.5.1. Consistency and agreement among raters is established in subsection 4.5.1 and then, the resulting cut-off scores in relation to the CEFR are investigated in subsection 4.5.2.

### **4.5.1 Methodology**

The standard setting method for GESE started with an explanation of the Manual variant of the Angoff method, which the judges felt was clear. The judges were asked to answer a modification of the question in the Manual (Council of Europe, 2003: 85): “At what CEFR level can a test taker already obtain each score?”. In order to write their answers, the judges were given forms such as the one in Appendix 6 (p. 89) and they were also asked to include comments should they wish. They were also given a handout containing CEFR scales (see also subsection 4.2.2).

Some judges felt that providing a level estimate for any possible score on the GESE suite was rather complicated, because of the large number of Grades (12); this meant that there were 36 possible scores overall, since each Grade would have three passing scores (pass, merit, distinction). As a result, five judges focused only on pass score, whereas six judges provided judgements for some of the other scores as well. All levels estimates were accompanied by justification, usually in the form of quoting relevant CEFR scales.

After the completion of the task, a group discussion about the standards set followed, with the estimate of the pass score by each judge shown on a screen using EXCEL. Following this discussion, the second round of judgements followed the question asked in the Extended Angoff Procedure (Hambleton & Plake, 1995), that is judges had to concentrate on the barely pass candidate. This caused some reactions, as the judges said that what they usually consider is a clear pass candidate between pass and merit. However, it was explained to them that in many contexts the barely pass person is very important to be defined, as such a person holds the same qualification as anyone who has passed the same Grade with a higher score, though Trinity certificates do confirm the level of pass attained.

After this round, the judges were also asked to state their level of confidence in the resulting cut-off score. In order to make sure the group did not influence individuals in stating confidence levels publicly, the feedback forms distributed after the end of the session contained a question on level confidence for the two suites. On a three-point scale (not at all confident, fairly confident and very confident), all judges felt very confident for the Initial GESE Grades and the ISE levels. Only 3 judges felt fairly confident for the Elementary, Intermediate and Advanced stages of GESE, but as they wrote, this did not have to do with the standard setting procedure, nor the test, but the correspondence between descriptors and the exam context. Positive comments pointed out that the three-day Standardisation meeting was very well organised and that the unanimity among the group members for the final cut-off scores leaves little room for error.

#### 4.5.2 Investigating consistency and agreement

Table 4.27 presents coefficients of consistency and agreement for the Standard Setting stage for GESE. Judgements are grouped into two rounds as can be seen in the first column. Judgements about the ISE cutscore are not analysed, because as I will show below, descriptive statistics show clearly very high levels of agreement. Overall, consistency and agreement are very high for the cut scores in relation to the CEFR.

*Table 4.27 Agreement and consistency of GESE cut-scores judgements*

Stage	Inter-rater reliability			Alpha	ICC**	W**
	Mean*	Min	Max			
Standard-setting 1 <sup>st</sup> round	0.986	0.946	1	0.998	0.998	0.991***
Standard-setting 2 <sup>nd</sup> round	0.996	0.98	1	0.999	0.998	0.993***

\* Average using Fisher's Z-transformation

\*\* Statistically significant at level  $p \leq 0.01$

\*\*\*Only scales used by all judges

#### 4.5.3 Cut-off scores in relation to the CEFR for GESE and ISE

In this section I will discuss the cut-off scores, as well as the judges' confidence in the standards set. Table 4.28 presents cut-off scores for GESE in the first round and Table 4.29 for the second round. As stated in subsection 4.5.1, the judges provided the estimated level for candidates receiving each of the scores in the GESE suite in the first round, whereas in the second round they only estimated the CEFR level of the borderline candidate for each GESE Grade.

Standard-setting methodology for ISE was identical to GESE. However, the judges did not provide, most probably because of tiredness, estimates of the CEFR levels for other scores apart from pass. Reliability statistics of judgements for ISE were not provided in subsection 4.5.2, because the very narrow range in Table 4.30 clearly demonstrates high agreement.

It is essential to clarify here that Claudia's judgements were included in establishing cut-off scores, even though she obtained a low score in some cases in the Familiarisation activities discussed in subsection 4.3.2; this decision was taken, because as can be seen in the descriptive statistics in the following Tables, measures of dispersion are very low, suggesting that the group has reached very high agreement and therefore Claudia does not appear to have behaved differently from the other judges.

Table 4.28 Cut-off scores in relation to the CEFR-GESE round 1

<b>GESE Grade</b>	<b>Trinity band score</b>	<b>N</b>	<b>Mean</b>	<b>Median</b>	<b>Mode</b>	<b>SD</b>	<b>Range</b>	<b>Min</b>	<b>Max</b>
<b>12</b>	<b>DISTINCTION</b>	2	22	22	22	0	0	22	22
	<b>MERIT</b>	2	22	22	22	0	0	22	22
	<b>PASS</b>	11	21.45	22	22	1.29	4	18	22
<b>11</b>	<b>DISTINCTION</b>	3	21.33	22	22	1.15	2	20	22
	<b>MERIT</b>	3	18.67	18	18	1.15	2	18	20
	<b>PASS</b>	10	18	18	18	0	0	18	18
<b>10</b>	<b>DISTINCTION</b>	4	17.5	18	18	1	2	16	18
	<b>MERIT</b>	4	17.5	18	18	1	2	16	18
	<b>PASS</b>	11	16.91	16	16	1.04	2	16	18
<b>9</b>	<b>DISTINCTION</b>	3	16.67	16	16	1.15	2	16	18
	<b>MERIT</b>	3	15.33	16	16	1.15	2	14	16
	<b>PASS</b>	11	15.09	16	16	1.04	2	14	16
<b>8</b>	<b>DISTINCTION</b>	3	14.67	14	14	1.15	2	14	16
	<b>MERIT</b>	3	14	14	14	0	0	14	14
	<b>PASS</b>	11	14	14	14	0	0	14	14
<b>7</b>	<b>DISTINCTION</b>	3	14	14	12	2	4	12	16
	<b>MERIT</b>	3	13.33	14	14	1.15	2	12	14
	<b>PASS</b>	11	12.82	12	12	0.98	2	12	14
<b>6</b>	<b>DISTINCTION</b>	4	12	12	12	0	0	12	12
	<b>MERIT</b>	4	12	12	12	0	0	12	12
	<b>PASS</b>	11	10.45	10	10	0.82	2	10	12
<b>5</b>	<b>DISTINCTION</b>	5	10	10	10	0	0	10	10
	<b>MERIT</b>	5	9.6	10	10	0.89	2	8	10
	<b>PASS</b>	11	9.46	10	10	1.57	6	6	12
<b>4</b>	<b>DISTINCTION</b>	5	7.6	8	8	0.89	2	6	8
	<b>MERIT</b>	5	7.6	8	8	0.89	2	6	8
	<b>PASS</b>	11	7.46	8	8	0.93	2	6	8
<b>3</b>	<b>DISTINCTION</b>	7	6.86	6	6	1.07	2	6	8
	<b>MERIT</b>	6	6	6	6	0	0	6	6
	<b>PASS</b>	11	5.82	6	6	0.6	2	4	6
<b>2</b>	<b>DISTINCTION</b>	7	4.57	6	6	1.9	4	2	6
	<b>MERIT</b>	6	3.33	3	2	1.63	4	2	6
	<b>PASS</b>	11	2.18	2	2	0.6	2	2	4
<b>1</b>	<b>DISTINCTION</b>	6	2	2	2	0	0	2	2
	<b>MERIT</b>	6	1.83	2	2	0.41	1	1	2
	<b>PASS</b>	11	1.63	2	2	0.5	1	1	2

Table 4.29 Cut-off scores in relation to the CEFR-GESE round 2

Grades	N	Mean	Median	Mode	SD	Range	Min	Max
Grade 12	11	20.73	20	20	1.35	4	18	22
Grade 11	11	18	18	18	0	0	18	18
Grade 10	11	16.91	16	16	1.04	2	16	18
Grade 9	11	15.09	16	16	1.04	2	14	16
Grade 8	11	14	14	14	0	0	14	14
Grade 7	11	12.73	12	12	1.01	2	12	14
Grade 6	11	10.36	10	10	0.81	2	10	12
Grade 5	11	9.46	10	10	0.93	2	8	10
Grade 4	11	7.64	8	8	0.81	2	6	8
Grade 3	11	5.82	6	6	0.6	2	4	6
Grade 2	11	2.55	2	2	0.93	2	2	4
Grade 1	11	1	1	1	0	0	1	1

Table 4.30 Cut-off scores in relation to the CEFR-ISE

ISE level	Round	N	Mean	Median	Mode	SD	Range	Min	Max
ISE III	1 <sup>st</sup>	11	18	18	18		0	18	18
	2 <sup>nd</sup>	11	18	18	18		0	18	18
ISE II	1 <sup>st</sup>	11	14	14	14		0	14	14
	2 <sup>nd</sup>	11	14	14	14		0	14	14
ISE I	1 <sup>st</sup>	11	10.18	10	10		2	10	12
	2 <sup>nd</sup>	11	10.18	10	10		2	10	12
ISE 0	1 <sup>st</sup>	11	6.91	6	6		2	6	8
	2 <sup>nd</sup>	11	7.09	8	8		2	6	8

#### 4.5.4 Discussion of the resulting cut-off scores

Based on the judgements presented in the previous subsection, linking GESE and ISE to the CEFR is possible now through Standardisation and a holistic judgement for each suite is provided in this subsection. It is essential to explain here how the holistic claim of the CEFR levels corresponding to specific Grades has been built

First of all, Table 4.31 presents the estimated level of the borderline person for GESE grades. This is an essential point in the standard setting process, not only because it is the aim of the standard setting methods used, but above all because it secures the validity of the linking claim. For example when a candidate passes Grade 7, such a certificate will probably be used for employment, studying in an English-speaking University and other important purposes. Even though there is some evidence in Table 4.28 that someone obtaining a distinction at Grade 7 could demonstrate a B2 performance, still a borderline pass candidate is a holder of the same certificate. For that reason, it is imperative that linking to the CEFR by Standardisation presents the minimum possible CEFR level, even if the actual borderline Grade 7 population is a small proportion of the overall population. In other words, even though we might know that only a small percentage of candidates belong to the borderline group, still any candidate belonging to this population will hold and use the same certificate as a distinction candidate. However, the judges suggested that the CEFR level of what they called the “secure pass candidate” should be indicated.

They felt that the performance of the secure pass candidate is more representative of the average candidate of each GESE Grade.

Following the above, Table 4.31 shows the minimum CEFR level expected for each Grade and the CEFR level of the secure pass candidate for each Grade. The former is arrived at by looking at the median of the judgements in the second round of Table 4.29. The latter is included in the third column of Table 4.31 and it is derived from the median of the B band score of each Grade in Table 4.28. Two things should be explained here about arriving to the holistic judgement: first, the use of the median and second, the reference to the second round of judgements. The median was chosen because it provides a more accurate picture of the estimated level since “it is not strongly affected by extreme scores” (Bachman, 2004:62). In the CEFR literature the median has been used by Alderson (2005:183, 201) in order to compare grammar and vocabulary descriptor difficulty as perceived by teachers and intended by the DIALANG levels. The second round of providing cut-off scores was preferred because judgements are expected to be even more valid than those in the first round, due to the group discussion between the two rounds. Table 4.28 provides some evidence from the first round of judgements that within each Grade a candidate’s performance could demonstrate characteristics of the higher band, that is, a merit or distinction at Grade 12 is expected to correspond to C2 as opposed to C1+ for a pass score. Finally, it should be pointed out that ‘+’ denotes that the performance expected at this level demonstrates stronger elements of the CEFR level, which is usually included in the upper part of some bands in a number of the CEFR scales.

*Table 4.31 CEFR level of borderline and secure pass candidates in the GESE suite*

<b>Grades</b>	<b>Level of borderline candidate</b>	<b>Level of secure pass candidate</b>
Grade 12	C1+	C2
Grade 11	C1	C1
Grade 10	B2+	C1
Grade 9	B2+	B2+
Grade 8	B2	B2+
Grade 7	B1+	B2
Grade 6	B1	B1+
Grade 5	B1	B1
Grade 4	A2+	A2+
Grade 3	A2	A2
Grade 2	A1	A1+
Grade 1	Below A1	A1

The minimum expected CEFR level for each ISE level is presented in Table 4.32 and is based on the same rationale as the GESE cut-offs. It might be the case that the high agreement indicated by measures of dispersion in Table 4.30 was because the judges were aware of the correspondence between the GESE Grades and ISE levels, which could have affected decision making as can be seen by comparing GESE Grades 4, 6, 8 and 11 with the four ISE levels. This is not unreasonable, since the Interview component in both suites is common. It should be also noted that the CEFR level estimate is a holistic one, since ISE levels comprise three components, Interview, Portfolio and Controlled Written Exam.

Table 4.32 CEFR level of borderline candidates in the ISE suite

ISE levels	Level of borderline candidate
ISE III	C1
ISE II	B2
ISE I	B1
ISE 0	A2+

#### 4.5.5 Considering the validity of the Standardisation claim

Kaftandjieva (2004:20-21) points out that there are two main types of validity evidence in standard setting : procedural and generalisability.

Procedural evidence (i.e. the suitability and proper implementation of the chosen standard setting procedures with regard to the concrete circumstances) cannot on its own guarantee the validity of cut-off scores interpretations, however lack of such evidence can affect negatively the credibility of the established cut-off scores (ibid). In order to secure procedural evidence, the following points, described in Hambleton (2001: 93) and Kaftandjieva (2004: 20) were considered in the present study:

- Selection of panellists: The rationale for selecting the panel has been explained in previous subsection 1.3, but the main intention was to work with people highly involved in various aspects of test construction and administration. For this reason the group consisted of panellists who were involved in examining, marking, item writing, monitoring, piloting and providing administration support. Bearing in mind that the selected standard setting methods would involve the definition of imaginary learners, an additional aim was that they were familiar with the candidature.
- Training of panellists: As has already been described in this report the judges followed a number of familiarisation activities to ensure their in-depth understanding of the CEFR and its scales, in order to apply it during the standard setting method. Evidence of judges' consistency and agreement has been in all sections of the report.
- Sequence of activities in the process: The procedure described in the Manual (Council of Europe, 2003), from Familiarisation to Training and Benchmarking, aimed at the building of common understanding of the CEFR scales on which the cut-off scores would be established, as well as the performance of Trinity candidates which would be judged according to the CEFR scales. It should be stressed that Familiarisation was carried out not only as a separate phase of the project, but also as part of Specification and Standardisation. In combination with the content description of the exams during Specification, it is reasonable to argue that the panel achieved a common understanding of the CEFR and the candidates whose performance was to be judged in relation to the CEFR.
- Documentation and feedback: Careful documentation was of primary concern, not only by taking notes and compiling reports, but also by recording the sessions in order to clarify points in the sessions that were not clear in the documentation. Feedback was also given during the Familiarisation session, in order to give judges a picture of group understanding of the CEFR scales.

A very simple way of increasing the precision of cut-off scores and therefore providing generalisability, the second type of validity evidence, is to increase the number of judges, items and occasions used in the standard setting (Kaftandjieva, 2004: 23). The number of judges used in the standard setting phase, as well as the

other phases, is within those suggested in the literature, which should raise the generalisability of the cut-off scores. Results of inter-rater and intra-rater consistency are quite positive, thus supporting generalisability evidence. The second round of cut-scores also aimed at increasing the occasions of providing cut-scores, whereas the large number of DVD samples, 16 in total, should be a satisfactory sample for illustrating Trinity candidates' performance, thus enhancing generalisability evidence.

Finally, as pointed out in subsection 4.5.1, judges' high confidence in the cut-scores they set should be included as another argument in favour of the validity of the resulting cut-off scores.

## **4.6 Conclusion**

Section 4 of this report has described in detailed the Standardisation phase of the project. The methodology was based on the Manual suggestions. Familiarisation Training, Benchmarking and Standard Setting sessions were organised and statistical analysis was employed to investigate consistency and agreement among judges and in addition to these, the cut-off scores in relation to the CEFR, which were accompanied by high levels of confidence by the judges.

## **5 Empirical Validation**

Chapter 6 of the Manual provides guidelines for Empirical Validation. The Manual points out that the claims of Specification and Standardisation need to be confirmed when test data become available (p. 100). Empirical validation is further divided for the purposes of a linking project into internal and external validation. Both of them will be addressed here, albeit not in great detail for reasons explained in each section separately.

### **5.1 Internal Validation**

Internal validation aims at answering a number of questions in relation to the quality of the exam. These questions are included in the Manual (pp. 100-101) and refer among others to individual items as well as the reliability of test scores awarded by raters and the extent to which examiners and raters follow their instructions during the assessment process. I will address these issues in the following subsections.

Internal validation should not only be viewed as an indispensable part of a CEFR linking project; it is also an on-going process which ensures that candidates receive valid, reliable and fair test scores. Given this and also the time-consuming nature of designing and conducting the previous three stages of the linking projects, it was clear from the beginning of piloting the Manual that a second researcher had to investigate internal validation. Alistair Van Moere of Lancaster University conducted a number of statistical analyses for the results of both GESE and ISE. Separate reports describe in detail these validation studies (Van Moere, 2006a, 2006b) which is why internal validation will only be discussed here briefly (subsections 5.1.1-5.1.3); for further details one can consult the corresponding reports.

Furthermore, subsections 5.1.4-5.1.11 will also discuss examiner training for GESE and ISE. Given that candidates' scores are based on ratings by examiners, the latter's training is crucial for securing reliability of test scores.

#### **5.1.1 The GESE study**

In the GESE study Van Moere (2006a), collected data from an 18 month period (May 2004 to September 2005) from 15 monitors who observed 142 examiners while administering GESE tests to 1,118 candidates in 132 different exam centres. It should be restated here that oral exams are a one-to-one interaction, with one candidate being interviewed and marked by the examiner. Examiners are equipped with a tape recorder and at least 10% of the interviews are audio recorded. As Van Moere explains (2006a:2) the data for the validation study consists of 'live monitoring' ratings which are part of the standards-maintenance procedures followed by Trinity. During monitoring a senior examiner (monitor) observes examiners' performances during live tests. Typically when senior examiners undertake a live monitoring they observe up to eight tests in one session. As well as giving feedback about personal conduct, questioning and language elicitation techniques, senior examiners separately award scores to candidates in the tests they monitor. A record is kept of both sets of ratings so that scores awarded by examiners and monitors can be compared and discussed. In the course of each year 33% of all Trinity GESE examiners are monitored.

Van Moere (2006a:2) notes that although it is not possible to establish ongoing reliability from single ratings, live monitoring data allows us to establish rater

reliability over an extended period of time for a large segment of the examiner panel. This is based on the following rationale:

[I]f it can be shown that different examiners are independently and consistently arriving at the same or similar scores for almost all candidates based on performances during live monitoring, then it would be reasonable to extrapolate from this that the remaining two-thirds of examiners in any given year are rating in a similarly reliable manner.

(Van Moere, 2006a:2)

The results for rater reliability were very encouraging. High level of agreement was observed between examiners and monitors for awarding Distinctions, Merits, Pass, and Fails at the overall results level. On average, examiners and monitors were in exact agreement on the final score to be awarded in over 70% of exams administered, and in 98% of exams they agreed exactly or agree to within one band, i.e. the difference between a Distinction and a Merit. This was consistent across all twelve Grades.

### 5.1.2 The ISE study

The ISE study (Van Moere, 2006b) dealt with candidates' scores on the Portfolio and the Controlled Written components. Portfolios that were monitored between December 2005 and June 2006 were first investigated. The data comprises double-markings that were assigned to 177 candidates' Portfolios by both an examiner and a monitor, from ISE Levels 0, I, II, and III. The Portfolios were marked once from a pool of 94 examiners and a second time from a pool of four monitors. All Portfolios consist of three tasks which are each graded according to Task Fulfilment criteria on a scale from A to E. The second data set consisted of double-markings that were assigned to candidates' Controlled Written exams from November 2005 to May 2006. Samples from 157 candidates representing the four ISE levels were marked once from a pool of six examiners and a second time from a pool of four monitors. The Controlled Written component consists of two tasks (ISE 0, I and II) and three tasks (ISE III), each also graded on a scale from A to E, according to two criteria, Task Fulfilment and Accuracy/Range.

All grades were converted into numbers so that grades on different components of the exam could be tallied and the candidate given a final overall score. Agreement coefficients were estimated using the grades assigned by the two raters whereas the numerical data were used for correlation coefficients. For the Controlled Written exam of ISE I Van Moere also applied many-facet Rasch analysis using FACETS, with three facets: examinees, raters and tasks.

Table 5.1 from Van Moere (2006b:4-7) shows the proportion of occurrences on which the examiner and the monitor either agreed exactly or agreed to within one band for the total score of candidates.

*Table 5.1 Examiners-monitors scoring agreement for ISE*

ISE levels	Agreement %-Portfolio	Agreement %- Controlled Written
ISE III	61	100
ISE II	77	97
ISE I	90	99
ISE 0	83	97

Agreement is high, especially for the Controlled Written component. ISE III Portfolio generated a lower result. Van Moere explains that this is the result of examiners and monitors disagreeing over B, C and D scores in ISE III more than in other ISE levels. Moreover, he explains that conversion of band scores into numbers for the purposes of statistical analysis magnifies the difference between awarded scores, even if for example one examiner awards C to all three portfolio tasks whereas a monitor awards D to the same tasks.

Rasch analysis showed that the two tasks in ISE I Controlled Written component were comparable in terms of difficulty, with Task Fulfilment for Task 2 slightly easier. The six raters involved in the marking of ISE I were operating in a similar way in terms of leniency/severity, with only one examiner being relatively lenient and assigning scores in a way that is different to the other raters; this was denoted by an infit mean square value of 1.70, whereas the other raters fell within the widely accepted .7-1.3 range. Finally, examinees were found to represent a wide range of abilities and performed relatively well in the exam; only a few of them displayed misfit.

A final consideration is expressed for the Portfolio component, similarly to the one made in subsection 4.4.9. Seven portfolios were given a 'U', that is unmarked, because of suspicion that the portfolio task was plagiarised. In five cases both the examiner and the monitor agreed on the unmarked choice, but in two cases one of the two assigned a score. Even though this is only a very small proportion out of the 177 portfolio examined in the study, it is worth investigating further to ensure that candidates are treated fairly.

### **5.1.3 Conclusion on the GESE and ISE studies**

The two studies discussed here (Van Moere, 2006a, 2006b) produced some positive results for the reliability of the procedures followed during the administration and marking of the exams. However, it should be stressed that Van Moere's work cannot on its own provide an exhaustive argument for internal validation. The two studies investigated mainly the reliability of test scores and as such, they contribute to one part of the validation process. As I have explained above, internal validation is an on-going process necessary for providing test scores of acceptable quality. Since examiner training is fundamental in contributing to reliability in performance assessment, the next subsections will deal with the training of the Trinity examiners.

### **5.1.4 Examiner training and its importance for the CEFR linking claim**

Training examiners to elicit candidates' language as interlocutors and assess it is a crucial component of any testing programme as many authors point out (Alderson et al., 1995: Ch. 5; Fulcher, 2003: Ch. 6; Luoma, 2004: Ch. 8). GESE and ISE are exams that fall in the category of what is usually called 'performance assessment' (McNamara, 1996). This means that candidates taking GESE or the spoken component of ISE will be engaged in a one-to-one interaction with an examiner who will then use marking criteria to assess their performance on a task. For the Portfolio component of ISE, written performance is assessed by the examiner who will then use the Portfolio to engage the candidate in discussion regarding the content of the written work. The Controlled Written component of ISE is marked centrally.

Following the above, I decided to examine the procedures followed by Trinity for training GESE and ISE examiners. This would offer Trinity an additional check of the procedures adopted for securing reliability of test scores, but in relation to the CEFR

project, monitoring examiner training was extremely important for the final claims in relation to the Framework, because linking to the CEFR is pointless if extra care is not taken to ensure reliable scores. In the case of GESE and ISE, given their performance assessment nature, examiner training is fundamental for score reliability.

### **5.1.5 Aims of examiner training for GESE and ISE**

It is clear from the previous subsection that training for GESE and ISE examiners needs to address the role of interlocutor, as well as assessor for oral performance and assessor of written performance for the ISE Portfolio. Therefore, monitoring the procedures for training examiners focused on how these points were addressed.

### **5.1.6 The Examiners' Conference**

The main training event for Trinity examiners takes place once a year under the name 'Examiners' Conference' and all examiners are required to attend. Examiners' Conference in January 2006 was attended by the author.

The event starts on Friday evening and finishes on Sunday afternoon in a large hotel in England, with accommodation provided for all examiners in the hotel. This means that all examiners receive exactly the same training on the same day, by the same group leaders. Arguably this is positive for score reliability, as it overcomes the practicality issue for large scale exams to provide centralised training. Trinity recruits UK-based examiners who then fly to countries where exams are administered instead of recruiting local examiners. The positive outcome of this is that all examiners are trained in a single event. 225 Trinity examiners participated in the 2006 Conference.

My intention during my attendance was to participate as if I were a novice examiner in order to judge whether the training event would cover the issues addressed in the previous subsection. Detailed notes were taken throughout the three days.

### **5.1.7 Description of the Conference programme-Day 1**

On the Friday, a plenary session by Trinity officials provided details on the activities of Trinity and more specifically the Language Group, which is responsible for GESE and ISE. I found this very informative, from a novice examiner's point of view. The last talk was dedicated to research conducted at Lancaster the year before (Shaw-Champion, 2005). Shaw-Champion conducted research during his attendance at the Language Testing at Lancaster (LTL) course, which is a summer course designed for those individuals or teams whose work or personal interests involve them in language test construction and development at the local or national level (<http://www.ling.lancs.ac.uk/study/languagetesting/lt2004.htm>). Trinity provides funding to examiners who want to attend the course and submit research related to Trinity exams.

Shaw-Champion's research investigated the Conversation phase of GESE Grade 6. He collected 10% of the monitoring tapes submitted in 2004 and the conversation phase of 21 interviews was fully transcribed. The aim of the study was to determine if there were different styles employed by the examiners and if these had effects on candidate performance.

Examiner behaviour was coded under six categories: Interrogator (direct questioning), Counsellor (mirroring, reflecting), Friend (empathising, informality, fun), Angler (eclectic, non-focussed), Teacher (eliciting target language) and Listener (non verbal encouragement). In order to quantify and report data, three measures were used: words spoken to indicate examiner and candidate talk, number of questions

asked to indicate degree of intervention and finally, functions and structures of the stage in order to suggest fulfilment of exam requirements.

After considering relevant research (Brown, 2003; Csepes & Egyud, 2004; Fulcher, 1996; Luoma, 2004), two approaches by examiners were identified. A Subject Interview approach was attributed to examiners who tended to follow conversation subjects of the syllabus without abrupt changes. A Grammar Interview approach was attributed to examiners tending to conduct a search for the grammar items specified in the syllabus. Shaw-Champion concluded: “[c]andidates in the Grammar interview, despite the high input by examiners, produced talk with less grammatical content; in the Subject interviews they produced more grammar items and used more than twice as much GESE Stage grammar as the examiner” (Shaw-Champion, 2005:5). Recommendations included inclusion of more standardised prompts in the documents given to examiners for conducting the conversation phase.

Shaw-Champion’s study was presented on Day 1 in order to familiarise examiners with its content and findings, which formed the basis for sessions on Day 2.

### **5.1.8 Description of the Conference programme-Day 2**

Day 2 started with the Chief Examiner ESOL summarising in a plenary session the aims of the Conference, that is, standardisation and reliability. Examiners were also invited to submit proposals for research as was the case with the study presented the day before. In the next session, one of the Academic Managers explained the procedures for quality assurance and described the monitoring system, which has also been mentioned in this report.

Three group sessions followed the plenary talk. Examiners are grouped into 14 teams assigned to an equal number of group leaders, with group sessions running in parallel for 90 minutes each. Session 1 focused on marking the ISE Portfolio, session 2 was about the requirements of the Conversation phase, and finally, session 3 drew on Shaw-Champion’s study, where the two identified approaches were discussed and illustrated with recorded samples. Examiners were invited to discuss with the other group member and their group leaders the two approaches, to identify advantages and disadvantages and suggest which approach is better.

The evening plenary sessions focused on operational issues, such as travelling to the country where exams are to be administered, procedures to be followed during the exam and suggestions for problem-solving during the exam. A final session discussed the issue of personal safety for examiners flying to other countries.

### **5.1.9 Description of the Conference programme-Day 3**

The Sunday sessions started with a plenary by the Chief Examiner, following the discussion of session 3 the previous day. The next two sessions involved assessing recorded GESE and ISE samples. This is similar to Benchmarking in section 4 of the present report and includes warm up, individual rating of samples and discussion of ratings. Rating forms are used and the ratings are taken by group leaders in order to be analysed at the Head Office and feedback to be posted to examiners after the Conference. The final session again by the Chief Examiner was a summary of the conference. It was pointed out that examiners need to ensure that they elicit the language of each Grade and ISE level specified in the syllabus, but at the same time the discussion should resemble authentic interaction as closely as possible.

### **5.1.10 Examiners' pack**

Examiners receive a pack containing a number of documents to be used when conducting the exams. The documentation was used during the Conference and comprises four booklets identified by different colours.

The Academic Procedures for GESE booklet is to be used along with the syllabus and provides guidelines on how to conduct the exam for all GESE stages. The Examination Materials booklet contains prompts for the Interactive tasks and texts for the Listening task (see also subsection 1.4). The Examiners' Handbook for ISE is a booklet providing guidelines for conducting the spoken component of ISE and marking the Portfolio. Finally, the Operational Procedures booklet provides guidance on travelling to the country where the exam is conducted and problem-solving tips during the exam.

This documentation is essential for examiners and appears to cover the vast majority of aspects of exam administration that are directly relevant to examiners. It is expected that this material should contribute positively to reliable scores and establishing such a positive impact could be the aim of future research. Given the findings of the study by Shaw-Champion, adding further points to the guidelines for conducting the Conversation phase could also be a topic for further research.

### **5.1.11 Conclusion on examiner training**

Training examiners is an essential part of an examination and because of the nature of the Trinity exams, attending the annual Examiners' Conference was of particular importance for linking GESE and ISE to the CEFR. Participating in the Conference activities as a novice examiner allowed for first-hand experience of the procedures followed by Trinity. Overall the Conference was well-organised, addressing the points raised in subsection 5.1.5. It is therefore expected that the Conference contributes positively to reliable scores and to a great extent Van Moere's research supports this claim empirically. Of course, further research could investigate the effect of training by comparing ratings from groups of examiners receiving different amounts of training. Moreover, it can be researched whether the observed agreement is the result of examiners' detailed understanding of guidelines and marking criteria and not just the result of being obliged to agree about the rating of a specific performance because of the ratings by other examiners in the Conference.

### **5.1.12 General conclusion on Internal Validation**

The CEFR project was combined with an external review of aspects related to reliability of test scores, as discussed in subsections 5.1.1- 5.1.11. It is widely accepted in the testing field after Messick's (1989) seminal paper on validity as unitary concept, that reliability is part of the validity of test scores. Therefore the previous subsections discussing reliability of test scores awarded by Trinity can be viewed as part of a validity argument for Trinity exams. However, since validation is an on-going process, the discussion in subsections 5.1.1- 5.1.11 should not be viewed as an exhaustive validation argument. The CEFR project contributes to further investigation of the quality of Trinity exams, but only on-going validation can ensure the quality of the exams constantly.

## **5.2 External Validation**

External validation in the Manual (p. 108) concerns an empirical confirmation of the cut-off scores set in relation to the CEFR. The fundamental consideration in designing an external validation study is the definition of a criterion measure (Council of Europe, 2003:113). According to the Manual external validation can be conducted through indirect linking or direct linking.

### **5.2.1 Indirect and direct linking: some considerations**

Indirect linking involves the definition of an anchor test already calibrated to the CEFR as the criterion. This is administered to a population to which the exam in question has also been administered. Linear regression or Item Response Theory can then be employed to investigate the relationship between the anchor/criterion and the exam. Unfortunately, practical reasons did not allow an indirect linkage study to be designed. Practical reasons had to do with time limitations, identifying an anchor test as well as a common test population to which both tests could be administered. This could be of course a suggestion for further research designed by Trinity. Another exam provider could be approached and different tests can be administered for the same population. An additional suggestion could also be administering the different Trinity exams to the same test takers and then investigate the relationship between the exams.

The lack of defining a criterion in the form of a test is addressed in the direct linking design through teachers' judgements (Council of Europe, 2003:116). The Manual points out that if a criterion is lacking, then a criterion has to be constructed, but not necessarily in the form of a test. Teachers' informed judgements of their students' performance can be used as the criterion in this case. However, as was the case with indirect linking, designing a study where teachers' judgements can be employed was practically impossible within the present project. It is, however, another possible suggestion for further research by Trinity in order to validate the linkage of the exam to the CEFR.

### **5.2.2 Addressing the issue of defining a criterion**

Despite the difficulties described in the previous subsection, defining an external criterion is attempted here. The GESE syllabus describes a candidate profile for each Grade using CEFR descriptors, based on an earlier, short-scale comparison of the Trinity syllabus and the CEFR scales (Davies, 2001), which was not, however, based on empirical data. These CEFR-based profiles could form an external criterion for investigating the linkage of GESE to the CEFR in the next subsection. The same approach is followed for ISE and is discussed in subsection 5.2.4. The ISE syllabus also includes CEFR-based profiles, which appear to be the outcome of the corresponding GESE Grades and research for the design of the ISE Portfolio and Controlled Written exams (Green, 2000), where reference to the CEFR levels is made.

### **5.2.3 Criterion-test comparison for GESE**

Table 5.2 summarises the CEFR level of each Grade, as well as the level of the borderline candidate as defined in the Standard Setting session in subsection 4.5.4. Figure 5.1, adapted from the Decision Table figure of the Manual (Council of Europe, 2003:121), presents a graphic crosstabulation of the relationship of the GESE syllabus profiles and the cut-off scores from the Standardisation phase of the project.

Table 5.2 CEFR level comparison for GESE

Grades	Syllabus	Level of borderline candidate
Grade 12	C2	C1+
Grade 11	C1	C1
Grade 10	C1	B2+
Grade 9	B2	B2+
Grade 8	B2	B2
Grade 7	B2	B1+
Grade 6	B1	B1
Grade 5	B1	B1
Grade 4	A2	A2+
Grade 3	A2	A2
Grade 2	A1	A1
Grade 1	A1	Below A1

Figure 5.1 offers a graphic representation of Table 5.2. The last column shows the total number of GESE Grades placed at the CEFR levels based on the syllabus. The inner box shows how Grades were spread according to the Standard Setting session of the present project, the total of which can be found in the last row. The shaded cells indicate same classification.

Figure 5.1 CEFR Decision Table for GESE

		Test (GESE standard setting)							
		Below A1	A1	A2	B1	B2	C1	C2	Total
Criterion (GESE Syllabus)	Below A1								0
	A1	1	1						2
	A2			2					2
	B1				2				2
	B2				1	2			3
	C1						1		2
	C2							1	1
	Total		1	1	2	3	3	2	0

In order to calculate a Spearman correlation coefficient I used the six CEFR level from Table 5.2 and converted them into numbers (A1=1, A2=2, etc). The Spearman correlation is .962. As the Manual suggests (p. 111), an index of agreement between the criterion and the test classification can be calculated by adding the shaded cells

indicating same classification. The index of agreement for GESE is 8/12 or 66.7%. The fact that the Grades in the non-shaded area are all located on the left of the diagonal grey cells also indicates that the Standard Setting session produced lower cut-off scores than the estimated level in the syllabus. This could be because of the standard setting method employed, according to which the borderline candidate should be defined; the GESE syllabus on the other hand does not define the borderline candidate but provides a general estimate of the intended level for each Grade.

#### 5.2.4 Criterion-test comparison for ISE

Table 5.3 and Figure 5.2 report on comparison between the intended CEFR level in the syllabus and the cut-off scores as a result of the Standard Setting session of the project.

Table 5.3 CEFR level comparison for ISE

ISE levels	Syllabus	Level of borderline candidate
ISE III	C1	C1
ISE II	B2	B2
ISE I	B1	B1
ISE 0	A2	A2+

Figure 5.2 CEFR Decision Table for ISE

		Test (GESE standard setting)							
		Below A1	A1	A2	B1	B2	C1	C2	Total
Criterion (ISE Syllabus)	Below A1								
	A1								
	A2			1					1
	B1				1				1
	B2					1			1
	C1						1		1
	C2							1	1
Total			1	1	1	1		4	

One can easily see that there is 100% exact agreement between the syllabus and the cut-off scores; obviously the Spearman correlation is 1, indicating absolute rank order agreement. It should be noted that calculation for exact agreement involved conversion of the six levels into numbers as in the previous subsection. If however, more than six CEFR levels were used for the numerical conversion, then A2 and A2+ in Table 5.3 would be differentiated and exact agreement would be 75%.

### **5.2.5 Conclusion on External Validation**

It should be stressed that the external validation discussed in this section is very simple and only indicative of the relationship between the intended level of the syllabus for each suite and the estimated level of the borderline candidate in the Standard Setting session of the CEFR project. It does not compare candidate performance with an anchor test and it does not involve teachers' judgements as suggested in the Manual. It also ignores the possibility of uneven candidate profiles (Council of Europe, 2003:75), that is, candidates whose performance is not at the same level for all skills or categories in the test.

Moreover, it should be clarified that even though the Manual suggests the criterion has priority over the test, this does not appear to be appropriate here, as the criterion was not chosen because of its 'proven' linkage to the CEFR (Council of Europe, 2003:111) but only as a way to compare the cut-off scores with the estimated level from the candidate profiles in the syllabus.

However, this comparison has also generated some useful results regarding linking to the CEFR and the number of levels offered by an exam. Results suggest that the greater the number of levels the more difficult it is to generate exact agreement between the criterion and the test.

### **5.3 Conclusion on Empirical Validation**

The Manual distinguishes between two kinds of empirical validation: internal and external. Internal validation has been briefly discussed here and Van Moere's work has been referenced, along with a description of the training offered to Trinity examiners. External validation has also been discussed, albeit confined to comparison between the syllabus for each exam and the cut-off scores in relation to the CEFR generated by the final phase of the present project. Agreement between the criterion and the exam was higher for ISE than GESE.

## 6 General Conclusion

The present report has built a claim about the relevance of the CEFR and the GESE and ISE suites administered by Trinity College London, following the methodology in the pilot version of the Manual for relating exams to the CEFR (Council of Europe, 2003). This claim is based on judgements by 12 panellists whose consistency and understanding of the CEFR scaled descriptors was examined throughout the duration of the project.

The results of the analysis of the panellists' consistency of ratings and their understanding of the CEFR were positive and were examined in all phases of the project (see section 2 and subsections 3.1 and 4.1). According to the Manual, this is essential for building a claim as to the relevance of an exam to the CEFR.

A first claim about the relevance of the exam to the CEFR has been made by analysing test content during Specification. The results of this phase were graphically presented in subsection 3.3 and show the CEFR level of the GESE Grades and ISE level according to the judgements made during Specification.

The next phase of the Manual linking process, Standardisation, has provided a set of cut-off scores in relation to the CEFR. These cut-off scores are presented and discussed in subsections 4.5.3 and 4.5.4. The resulting cut-off scores have also been examined in relation to the syllabus candidate profiles in the Empirical Validation section (subsection 5.2).

The results of this project may be of interest to a number of parties, as explained in the Introduction. However, it should be stressed that the value of a linking project is only retained with continuous validation of test scores and frequent standardisation to ensure that the resulting cut-off scores in relation to the CEFR are valid across different versions of the exam. For this reason the detailed documentation of the present project is also useful as a set of guidelines for further research on standardising Trinity test scores on the levels of the Common European Framework of Reference.

A final word has to do with the pass score set by Trinity and the interpretation of the cut-off scores in relation to the CEFR in this study. In the GESE and ISE exams candidates receive a compensatory composite score (Bachman, 2004:318). For example a Grade 8 candidate might fail the Topic Presentation but pass the Interactive and Conversation phases. In this case the overall score will be Pass; details can be found in the syllabus for each exam. The cut-off scores set in this study are holistic and correspond to this overall, final score. Therefore, for the Grade 8 example, results suggest that the overall performance matches B2 level and not necessarily that B2 is the case for all phases of the exam.

## 7 References

- Alderson, J. C. (2002a). Using the Common European Framework in language teaching and assessment. In J. C. Alderson (Ed.), *Common European Framework of Reference for Languages: Learning, Teaching, Assessment: Case Studies* (pp. 1-8). Strasbourg: Council of Europe.
- Alderson, J. C. (2005). *Diagnosing foreign language proficiency: the interface between learning and assessment*. London: Continuum.
- Alderson, J. C. (Ed.). (2002b). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Case Studies*. Strasbourg: Council of Europe.
- Alderson, J. C., & Banerjee, J. (2001). Language testing and assessment (Part 1). *Language Teaching*, 34(4), 213-236.
- Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.
- Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2006). Analysing Tests of Reading and Listening in Relation to the Common European Framework of Reference: The Experience of the Dutch CEFR Construct Project. *Language Assessment Quarterly*, 3(1), 3-30.
- Andrich, D. (1978). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement*, 38(3), 665-680.
- Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.
- Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: fundamental measurement in the human sciences*. Mahwah, N.J.: Lawrence Erlbaum Associates.
- Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, 20(1), 1-25.
- Callahan, S. (1995). Portfolio expectations: possibilities and limits. *Assessing writing*, 2(2), 117-151.
- Carlsen, C., & Moe, E. (2005). *Basing writing tests for young learners to the CEFR*. Paper presented at the Second Annual Conference of EALTA, Voss, Norway, 2nd -5th June 2005. Retrieved 13/09/2005, from <http://www.ealta.eu.org/conference/2005/handouts/Moe%20and%20Carlsen/Voss%20presentation.ppt> .
- Cohen, L., Manion, L., & Morrison, K. (2000). *Research Methods in Education* (5th ed.). London: Routledge.
- 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: Manual, Preliminary Pilot Version*. Strasbourg: Council of Europe.

- Csepes, I., & Egyud, G. (2004). *Into Europe: The Speaking handbook*. Budapest: British Council.
- Davidson, F., & Lynch, B. K. (2002). *Testcraft*. New Haven: Yale University Press.
- Davies, S. (2001). *GESE initial benchmarking against the Common European Framework of Reference for Languages*. London: Trinity College London.
- Figueras, N., North, B., Takala, S., Verhelst, N., & Van Avermaet, P. (2005). Relating examinations to the Common European Framework: a manual. *Language Testing*, 22(3), 261–279.
- Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, 13(2), 208-238.
- Fulcher, G. (2003). *Testing second language speaking*. Harlow: Pearson Education.
- Generalitat de Catalunya. (2006). *Proficiency scales: the Common European Framework of Reference for Languages in the Escoles Oficials d'Idiomes in Catalunya*. Madrid: Cambridge University Press.
- Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement*, 15(4), 237-261.
- Gottlieb, M. (1995). Nurturing student learning through portfolios. *TESOL Journal*, 5(1), 12-14.
- Green, R. (2000). *Integrated Skills Exam (Report on mini-pilot)*. London: Trinity College London.
- Hambleton, R. (2001). Setting performance standards on Educational Assessments and Criteria for Evaluating the Process. In G. J. Cizek (Ed.), *Setting performance standards : concepts, methods, and perspectives*. Mahwah, N.J. ; London: Lawrence Erlbaum Associates.
- Hambleton, R., & Plake, B. (1995). Using an Extended Angoff Procedure to Set Standards on Complex Performance Assessments. *Applied Measurement in Education*, 8(1), 41-55.
- Hambleton, R., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park ; London: Sage Publications.
- Hamp-Lyons, L., & Condon, W. (1993). Questioning the assumptions about portfolio-based assessment. *College Composition and Communication*, 44(2), 176-190.
- Hamp-Lyons, L., & Condon, W. (1999). *Assessing the portfolio : principles for practice, theory, and research*. Cresskill, N.J.: Hampton Press.
- Hasselgreen, A. (2005). Assessing the language of young learners. *Language Testing*, 22(3), 337-354.
- Kaftandjieva, F. (2004). *Standard Setting. Section B of the Reference Supplement to the Preliminary Version of the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: learning, teaching and assessment*. Strasbourg: Council of Europe.
- Kaftandjieva, F., & Takala, S. (2002). Council of Europe Scales of Language Proficiency: A Validation Study. In J. C. Alderson (Ed.), *Common European Framework of Reference for*

- Languages: Learning, Teaching, Assessment. Case Studies.* (pp. 106-129). Strasbourg: Council of Europe.
- Klenowski, V. (2000). Portfolios: promoting teaching. *Assessment in Education*, 7(2), 215-236.
- Linacre, J. M. (1989). *Many-Facet Rasch Measurement*. Chicago: MESA Press.
- Linacre, J. M. (2005). Facets Rasch measurement computer program version 3.58. Chicago: Winsteps.com.
- Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.
- Martyniuk, W. (2006). *Relating language examinations to the Common European Framework of Reference for Languages (CEFR)*. Paper presented at the Into Europe-European Standards in Language Assessment Conference, Budapest, Hungary, 9th-10th February 2006. Retrieved 20/9/2006, from [http://www.examsreform.hu/Media/Relatinglanguage\\_exam.ppt](http://www.examsreform.hu/Media/Relatinglanguage_exam.ppt).
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174.
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1), 30-46.
- McNamara, T. (1996). *Measuring second language performance*. Harlow: Longman.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: Macmillan.
- Newman, C., & Smolen, L. (1991). Portfolio assessment in our schools: implementation, advantages and concerns. *Mid-Western Educational Researcher*, 6, 28-32.
- Nichols, D. P. (2006). *Choosing an intraclass correlation coefficient*. Retrieved 26/09/2006, from <http://www.utexas.edu/its/rc/answers/spss/spss4.html>.
- North, B. (2004). Relating assessments, examinations and courses to the CEF. In K. Morrow (Ed.), *Insights from the Common European Framework* (pp. 77-90). Oxford: Oxford University Press.
- North, B., & Hughes, G. (2003). *CEF Performance Samples for Relating Language examinations to the Common European Framework of Reference for Languages: Learning Teaching and Assessment.*, from <http://www.coe.int/T/DG4/Portfolio/documents/videoperform.pdf>
- Papageorgiou, S. (2006). *The use of qualitative methods in relating exams to the Common European Framework: What can we learn?* Paper presented at the Third Annual Conference of EALTA, Krakow, Poland. Retrieved 7/6/2006, from [http://www.ealta.eu.org/conference/2006/docs/Papageorgiou\\_ealta2006.ppt](http://www.ealta.eu.org/conference/2006/docs/Papageorgiou_ealta2006.ppt).
- Popham, W. J. (1978). As always, provocative. *Journal of Educational Measurement*, 15(4), 297-300.
- Salazar, A. J. (1995). Understanding the synergistic effects of communication in small groups: Making the most out of group member abilities. *Small Group Research*, 26(2), 169-199.

- Shaw-Champion, G. (2005). *Grade 6 Conversation phase: Examiner behaviour and candidate performance*. (Research report). Lancaster University, UK: Department of Linguistics and English Language.
- Tanko, G. (2004). *Into Europe: The Writing handbook*. Budapest: British Council.
- Trinity College London. (2005a). *Graded examinations in spoken English 2004-2007*. (2nd impression). London: Trinity College London.
- Trinity College London. (2005b). *Integrated Skills in English examinations 0, I, II, III* (4th ed.). London: Trinity College London.
- Tsagari, C. (2005). Portfolio assessment with EFL young learners in Greek State schools. In P. Pavlou & K. Smith (Eds.), *Serving TEA to Young Learners: Proceedings of the Conference on Testing Young Learners organized by the University of Cyprus, IATEFL and CyTEA* (pp. 74-90). Nicosia: ORANIM – Academic College of Education.
- van Ek, J. A., & Trim, J. L. M. (1998). *Threshold 1990*. Cambridge: Cambridge University Press.
- Van Moere, A. (2006a). *Graded Examination in Spoken English (GESE)* (External Validation Report). Lancaster University, UK: Department of Linguistics and English Language.
- Van Moere, A. (2006b). *Integrated Skills in English Examinations (ISE)* (External Validation Report). Lancaster University, UK: Department of Linguistics and English Language.
- Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15(2), 263-287.

## **Appendix 1: Familiarisation programme**

### **Programme of the Familiarisation meeting, London, 6-7 September 2005.**

#### **Tuesday, 6 September 2005**

9.30-10.00	Reception
10.00-11.15	Presentation of the Common European Framework and the Trinity calibration project by Spiros Papageorgiou, Project Coordinator.
11.15-11.30	Coffee break
11.30-13.00	Plenary discussion: Chapter 3 of the Common European Framework volume.
13.00-14.00	Lunch break
14.00-15.30	Familiarisation task 1
15.30-16.00	Coffee break
16.00-17.30	Familiarisation task 2
17.30-18.00	Plenary discussion: The CEFR descriptors

#### **Wednesday, 7 September 2005**

9.00-10.30	Familiarisation task 3
10.30-11.00	Coffee break
11.30-13.00	Familiarisation task 4
13.00-14.00	Lunch break
14.00-15.30	Familiarisation task 5
15.30-15.45	Break
15.45-16.30	Familiarisation task 6
16.30-17.00	Plenary discussion: The CEFR descriptors and the next phase of the project

## **Appendix 2: Specification programme**

### **Programme of the Specification meeting, London, 22-24 November 2005**

#### **Tuesday, 22 November 2005**

09.30-10.00	Reception
10.00-11.30	Summary of findings from the Familiarisation stage Familiarisation task for the group
11.30-11.45	Coffee break
11.45-13.00	Introduction to the Specification stage and discussion of instruments to be used
13.00-13.45	Lunch
13.45-15.15	Specification session 1: GESE Initial Grades-group work
15.15-15.30	Coffee break
15.30-16.30	Specification session 2: GESE Initial Grades-group work
16.30-16.45	Coffee break
16.45-18.00	Specification session 3: GESE Initial Grades-plenary discussion

#### **Wednesday, 23 November 2005**

09.15-10.30	Specification session 4: GESE Elementary Grades-group work and plenary
10.30-10.45	Coffee break
10.45-12.00	Specification session 5: GESE Elementary Grades-plenary and GESE Intermediate Group work
12.00-12.15	Coffee break
12.15-13.30	Specification session 6: GESE Intermediate Grades-plenary discussion
13.30-14.15	Lunch break
14.15-15.30	Specification session 7: GESE Advanced Grades-group work
15.30-15.45	Coffee break
15.45-17.00	Specification session 8: GESE Advanced Grades-plenary discussion

#### **Thursday, 24 November 2005**

09.15-10.30	Specification session 9: ISE 0, I, II and III-group work
10.30-10.45	Coffee break
10.45-12.00	Specification session 10: ISE 0, I, II and III - group work
12.00-12.15	Coffee break
12.15-13.30	Specification session 11: ISE 0, I, II and III - plenary discussion
13.30-14.15	Lunch break
14.15-15.30	Specification session 12: refining levels among GESE Grades
15.30-15.45	Coffee break
15.45-17.00	Summary of Specification stage and introduction to the Standardisation stage

## **Appendix 3: Standardisation programme**

### **Programme of the Standardisation meeting, London, 28 February-2 March 2006**

#### **Tuesday, 28 February 2006**

09.30-10.00	Reception
	Summary of findings from the Specification stage
10.00-11.30	Introduction to the Standardisation stage: aims and process
	Familiarisation task for the group
11.30-11.45	Coffee break
11.45-13.00	Familiarisation task: Discussion
13.00-13.45	Lunch
13.45-15.00	Standardisation session 1: Training with Calibrated samples
15.00-15.15	Coffee break
15.15-16.30	Standardisation session 2: Training with Calibrated samples
16.30-16.45	Coffee break
16.45-18.00	Standardisation session 3: Training with Calibrated samples

#### **Wednesday, 1 March 2006**

09.15-10.30	Standardisation session 4: Benchmarking-GESE Initial Grades
10.30-10.45	Coffee break
10.45-12.00	Standardisation session 5: Benchmarking-GESE Elementary Grades
12.00-12.15	Coffee break
12.15-13.30	Standardisation session 6: Benchmarking-GESE Intermediate Grades-
13.30-14.15	Lunch break
14.15-15.30	Standardisation session 7: Benchmarking-GESE Advanced Grades-
15.30-15.45	Coffee break
15.45-17.00	Standardisation session 8: Benchmarking-ISE 0 and I

#### **Thursday, 2 March 2006**

09.15-10.30	Standardisation session 9: Benchmarking-ISE II and III
10.30-10.45	Coffee break
10.45-12.00	Standardisation session 10: Standard setting: GESE Initial and Elementary
12.00-12.15	Coffee break
12.15-13.30	Standardisation session 11: Standard setting: GESE Intermediate and Advanced
13.30-14.15	Lunch break
14.15-15.30	Standardisation session 12: ISE-all levels
15.30-15.45	Coffee break
15.45-17.00	Summary of Standardisation stage

**Appendix 4: Samples from the two types of Familiarisation tasks**



**CEFR CALIBRATION PROJECT**

**Material from the Common European Framework for the Familiarisation phase**

**London, 6-7 September 2005**

**Name of project participant: .....**

**Guidelines:**

- 1) **Indicate the date on the top right side of the first page of each set of descriptors.**
- 2) **Please write on the right side of each descriptor the CEFR level (A1-C2) that you believe it belongs to. Choose only ONE level.**
- 3) **While doing the task, underline the words of the descriptor that you think are the key words that helped you decide on the level of the descriptor.**
- 4) **Feel free to write any comments you want for each descriptor, especially parts of the statements that troubled you.**
- 5) **After you finish the task, write on the left side of the statements the correct level of each descriptor, which will be given to you. DO NOT ERASE THE LEVEL YOU CHOSE AT THE BEGINNING, WRITTEN ON THE RIGHT SIDE.**

**Common Reference Levels: self-assessment grid**  
**Speaking**

---

**S1**

I can express myself fluently and convey finer shades of meaning precisely.

---

**S2**

I can connect phrases in a simple way in order to describe events.

---

**S3**

I can use simple phrases to describe where I live.

---

**S4**

I can use a series of phrases and sentences to describe in simple terms my family and other people.

---

**S5**

I can use language flexibly and effectively for social purposes.

---

**S6**

I can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible.

---

**S7**

I can present clear, detailed descriptions on a wide range of subjects related to my field of interest.

---

**S8**

I can narrate a story and describe my reactions.

---

**S9**

I can use language flexibly and effectively for professional purposes.

---

**Common Reference Levels: qualitative aspects of spoken language use**

Guidelines: Fill in the cells with the descriptors that will be given to you. A suggestion is that you first identify the descriptors that belong to each category (range, accuracy, etc.) and then decide on the level.

	<b>RANGE</b>	<b>ACCURACY</b>	<b>FLUENCY</b>	<b>INTERACTION</b>	<b>COHERENCE</b>
<b>C2</b>					
<b>C1</b>					
<b>B2</b>					
<b>B1</b>					
<b>A2</b>					
<b>A1</b>					

## Appendix 5: Rating Form for Speaking

### CEFR Rating Form for Speaking

DETAILS	
Your name:	
Learner's name:	

LEVEL ASSIGNMENT USING SCALED DESCRIPTORS FROM THE CEFR					
RANGE	ACCURACY	FLUENCY	INTERACTION	COHERENCE	OVERALL

Justification/rationale (Please include reference to documentation)

(Continue overleaf if necessary)

## Appendix 6: Standard Setting form for the Initial Grades

Please try to answer the following question:

**At what CEFR level can a test taker already obtain each score?**

In order to answer the question you might find useful to visualise learners you have met and examined, or consider the Trinity videos we used for benchmarking.

Write your decision on the third column, making sure that you have justified your decision by making reference to the CEFR scales.

GESE GRADE	SCORE	CEFR LEVEL OF THE CANDIDATE AND JUSTIFICATION
Grade 3	A	
	B	
	C	
Grade 2	A	
	B	
	C	
Grade 1	A	
	B	
	C	

## Appendix 7: Ratings of samples using Trinity bands

Examiner	Grade 1		Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10	Grade 12
	Erik	Carla	Christina	Marco	Susana	Monica	Renate	Michele	Paolo	Christian	Gabriel	Bobì
<b>Kate</b>	A	B	B	A	BC	BC	BB	CCC	CBC	DCC	BBBDB	BAAAA
<b>Andrew</b>	A	B	B	A	C	B	BB	CBB	CCC	CCC	CBBDC	BAAAA
<b>Matt</b>	A	B	B	A	AB	CC	AB	BBB	BBB	DCD	BBBDB	BAAAA
<b>George</b>	A	B	B	A	BC	CC	CB	BCB	CBB	DCD	DCBCC	AAAA
<b>Claudia</b>	A	B	B	A	BB	BB	BB	BCB	CBC	CBC	CBADB	AAAA
<b>Alice</b>	A	B	B	A	BC	BC	BB	CCB	BCC	DCD	CBBDD	BBAAA
<b>Tim</b>	A	A	B	A	BC	BC	CB	CBB	CCC	DDD	CBADB	AAAAA
<b>Roseanne</b>	A	B	B	A	BC	BC	AB	BBC	CBC	DCD	CBADB	AAAAA
<b>Rita</b>	A	B	B	A	BC	BC	BB	BBC	CBC	DCD	CBBDB	BAAAA
<b>Lora</b>	A	B	B	A	BC	BC	AA	BBB	CBC	CBC	BBCDC	BABAA
<b>Sally</b>	A	B	A	A	BB	BB	BB	BCB	CBC	DCD	CBADB	AAAAA

Examiner	ISE 0	ISE I	ISE II	ISE III
	Monica	Virginie	Mathilde	Atahan
Kate	BB	CC	CBC	CCBCB
Andrew	BB	CC	CBC	BBCCC
Matt	AA	BB	CCB	CCCCC
George	AA	BB	BBC	BCCCC
Claudia	BB	CC	CBC	BBCCB
Alice	BA	CC	CBB	BBCCC
Tim	AB	DC	CBB	CBCBB
Roseanne	BB	CC	CBB	CBACB
Rita	BB	CC	CBC	BBCCB
Lora	AA	CB	CBC	CCACB
Sally	BB	CC	CBB	BBCCB

Score	ISE 0		ISE I		ISE II		ISE III	
	CW*	Port.	CW*	Port.	CW*	Port.	CW*	Port.
<b>Score</b>	BCBC	ABB	CBCC	BBC	BBBB	ABA	CBABBB	CCB

\*The Controlled Written exam has two tasks for all levels apart from ISE III, which has three tasks. All tasks receive a score for Task Fulfilment and a score for Accuracy/Range