TRINITY
COLLEGE LONDON

Relating the Trinity College London GESE and ISE examinations to the Common European Framework of Reference

Final Project Report, February 2007

Spiros Papageorgiou
Department of Linguistics and English Language, Lancaster University

Trinity College London
89 Albert Embankment
London SE1 7TP UK

T +44 (0)20 7820 6100
F +44 (0)20 7820 6161
E info@trinitycollege.co.uk
www.trinitycollege.co.uk

Patron HRH The Duke of Kent KG

Published by Trinity College London

Contents

List of Tables 6

List of Figures 7

Foreword 8

Acknowledgements 13

1. Introduction 14

1.1 Aims of the project and outline of methodology 14

1.2 Significance of the project 15

1.3 Selection of participants 15

1.4 The GESE and ISE suites 16

1.5 Structure of the present report 16

2. Familiarisation 17

2.1 Methodology 17

2.1.1 Before the project meeting 17

2.1.2 During the project meeting 17

2.1.3 After the project meeting 18

2.2 Analysis of judgements 18

2.2.1 Analysis of judgements using classical statistics 18

2.2.2 Analysis of judgements using the Rasch model 20

2.2.3 Analysis of judgements from CEFR Table 3 and Manual 5.8 24

2.3 Conclusion 25

3. Specification 26

3.1 Familiarisation tasks 26

3.1.1 Analysis of judgements using classical statistics 26

3.1.2 Analysis of judgements using the Rasch model 28

3.1.3 Conclusion 29
3.2 Methodology 29
3.2.1 Before the meeting 29
3.2.2 During the meeting 30
3.2.3 After the meeting 31
3.3 Results 32
3.4 Discussion of methodology and results 34
3.4.1 The branching approach 34
3.4.2 Fit between the CEFR descriptors and the Trinity suite 34
3.4.3 Justification and rationale behind decisions 35
3.4.4 The link between GESE and ISE 35
3.4.5 The skill of Listening 35
3.4.6 Using the Specification Forms in practice 35
3.4.7 Validity of the Specification claim 35
3.5 Conclusion 36
4. Standardisation 37
4.1 Familiarisation tasks 37
4.1.1 Analysis of judgements using classical statistics 37
4.1.2 Analysis of judgements using the Rasch model 39
4.1.3 Conclusion 40
4.2 Methodology 40
4.2.1 Before the meeting 40
4.2.2 During the meeting 41
4.2.3 After the meeting 42
4.3 Training 43
4.3.1 Investigating consistency and agreement 43
4.3.2 Investigating the rating process 44

4.3.3 Conclusion 45

4.4 Benchmarking 45

4.4.1 Investigating consistency and agreement for the GESE suite 45

4.4.2 Investigating the rating process for GESE Initial Grades 46

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 54

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

4.5 Standard-setting 60

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

4.6 Conclusion 65

5. Empirical Validation 66

5.1 Internal Validation 66

5.1.1 The GESE study 66

5.1.2 The ISE study 66

5.1.3 Conclusion on the GESE and ISE studies 67

5.1.4 Examiner training and its importance for the CEFR linking claim 68

5.1.5 Aims of examiner training for GESE and ISE 68
5.1.6 The Examiners' Conference 68
5.1.7 Description of the Conference programme – Day 1 68
5.1.8 Description of the Conference programme – Day 2 69
5.1.9 Description of the Conference programme – Day 3 69
5.1.10 Examiners' pack 69
5.1.11 Conclusion on examiner training 70
5.1.12 General conclusion on Internal Validation 70
5.2 External Validation 70
5.2.1 Indirect and direct linking: some considerations 70
5.2.2 Addressing the issue of defining a criterion 71
5.2.3 Criterion-test comparison for GESE 71
5.2.4 Criterion-test comparison for ISE 72
5.2.5 Conclusion on External Validation 72
5.3 Conclusion on Empirical Validation 73
6. General Conclusion 74
7. References 75
Appendix 1: Familiarisation programme 78
Appendix 2: Specification programme 79
Appendix 3: Standardisation programme 80
Appendix 4: Samples from the two types of Familiarisation tasks 81
Appendix 5: Rating Form for Speaking 84
Appendix 6: Standard-setting form for the Initial Grades 85
Appendix 7: Ratings of samples using Trinity bands 86
Trinity College London response to the Final Report on the project to link GESE and ISE examinations to the CEFR 87

List of Tables

Table 1. The structure of the GESE suite 16
Table 2. The structure of the ISE suite 16
Table 2.1 Intra-rater reliability – Summary statistics 18
Table 2.2 Inter-rater reliability and internal consistency – Summary statistics 19
Table 2.3 Rater-CEFR agreement – Summary statistics 19
Table 2.4 Scores obtained by judges in the familiarisation tasks – Correct answers 20
Table 2.5 Rater measurement report – Familiarisation 21
Table 2.6 Scaling of the descriptors – Round 1 23
Table 2.7 Scaling of the descriptors – Round 2 24
Table 2.8 Correct placement of descriptors for CEFR Table 3 and Manual Table 5.8 24
Table 3.1 Intra-rater reliability – Summary statistics 26
Table 3.2 Inter-rater reliability and internal consistency – Summary statistics 27
Table 3.3 Rater-CEFR agreement – Summary statistics 27
Table 3.4 Scores obtained by judges in the familiarisation tasks – Correct answers 27
Table 3.5 Judges' characteristics – Logits and infit mean square values 28
Table 3.6 Scaling of the descriptors 29
Table 3.7 Organisation of the parallel sessions for GESE and ISE 30
Table 3.8 Rounds of judgements during the Specification stage 31
Table 3.9 Relationship of GESE content to the CEFR – Holistic estimation 34
Table 3.10 Relationship of ISE content to the CEFR – Holistic estimation 34
Table 4.1 Intra-rater reliability – Summary statistics 37
Table 4.2 Inter-rater reliability and internal consistency – Summary statistics 38
Table 4.3 Rater-CEFR agreement – Summary statistics 38
Table 4.4 Scores obtained by judges in the familiarisation tasks – Correct answers 38
Table 4.5 Judges' characteristics – Logits and Infit mean square values 39
Table 4.6 Scaling of the descriptors 40
Table 4.7 Conversion of CEFR levels into quantitative data 43
Table 4.8 Agreement and consistency of judges – Training sessions 44
Table 4.9 Training results – Summary statistics 44
Table 4.10 Items from the Matriculation Examination in Finland 44
Table 4.11 Agreement and consistency of judges – GESE benchmarking 45
Table 4.12 Benchmarking Initial Grades – Summary statistics 46
Table 4.13 Benchmarking Elementary Grades – Summary statistics 47
Table 4.14 Benchmarking Intermediate Grades – Summary statistics 48
Table 4.15 Benchmarking Advanced Grades – Summary statistics 50
Table 4.16 Agreement and consistency of judges – ISE Interview benchmarking 50
Table 4.17 Benchmarking ISE 0 Interview – Summary statistics 51
Table 4.18 Benchmarking ISE I Interview – Summary statistics 52
Table 4.19 Benchmarking ISE II Interview – Summary statistics 53
Table 4.20 Benchmarking ISE III Interview – Summary statistics 53
Table 4.21 Agreement and consistency of judges – ISE I and II Written benchmarking 54
Table 4.22 Benchmarking ISE I Written – Summary statistics 55
Table 4.23 Benchmarking ISE II Written – Summary statistics 56
Table 4.24 Agreement and consistency of judges – ISE 0 and III Written benchmarking sessions 57
Table 4.25 Benchmarking ISE 0 Written – Summary statistics 58
Table 4.26 Benchmarking ISE III Written – Summary statistics 59
Table 4.27 Agreement and consistency of GESE cut-scores judgements 61

Table 4.28 Cut-off scores in relation to the CEFR – GESE round 1 62
Table 4.29 Cut-off scores in relation to the CEFR – GESE round 2 63
Table 4.30 Cut-off scores in relation to the CEFR – ISE 63
Table 4.31 CEFR level of borderline and secure pass candidates in the GESE suite 64
Table 4.32 CEFR level of borderline candidates in the ISE suite 64
Table 5.1 Examiners-monitors scoring agreement for ISE 67
Table 5.2 CEFR level comparison for GESE 71
Table 5.3 CEFR level comparison for ISE 72

List of Figures

Figure 2.1 Ruler map for the Speaking 1 task 21
Figure 3.1 Graphic Profile of the GESE-CEFR relationship (Initial and Activities) 32
Figure 3.2 Graphic Profile of the GESE-CEFR relationship (Competences) 32
Figure 3.3 Graphic Profile of the ISE-CEFR relationship (Initial and Activities) 33
Figure 3.4 Graphic Profile of the ISE-CEFR relationship (Competences) 33
Figure 5.1 CEFR Decision Table for GESE 71
Figure 5.2 CEFR Decision Table for ISE 72

Foreword

Trinity College London's ESOL (English for Speakers of Other Languages) examinations are unique in the world of English language assessment. The following report details the research project carried out to calibrate Trinity's examinations to the *Common European Framework of Reference for Languages: Learning, teaching, assessment* (CEFR). Readers of the report should be aware of the format and content of Trinity's examinations and how they differ from other ESOL examinations as this will give insight into Trinity's experience of the calibration process.

Internationally, Trinity administers two suites of examinations. The Graded Examinations in Spoken English (GESE) assess skills in speaking and listening whereas the Integrated Skills in English examinations (ISE) assess the four skills of speaking, listening, reading and writing. Both suites, however, focus on real-life skills and communication. They are unique in their format, history, underlying philosophy and impact.

Unique formats

The format of the Graded Examinations in Spoken English (GESE) is simple: each examination consists of a one-to-one, unscripted, face-to-face interview with a trained and standardised native speaker examiner who does not live in the candidate's home country.

There are twelve grades in the GESE suite — these cover the whole range of English language competence, from pre-A1 on the CEFR (Grade 1) to C2 level (Grade 12). The twelve grades are grouped into four stages (Initial, Elementary, Intermediate and Advanced); at each stage the examination includes a new task or tasks and the length of the interview increases from 5 minutes (Grade 1) to 25 minutes (Grade 12).

The tasks are as follows:

General conversation	Included in all grades, this is an unscripted conversation designed to allow candidates to demonstrate their ability to use the language specified for the relevant grade in a real interaction with the examiner. The examiner assesses the candidate's contributions in terms of task fulfilment. Success does not rely wholly on the candidate's knowledge of individual words or phrases, but on their ability to use English to communicate successfully about conversation areas which are listed in the syllabus.
Topic discussion	Included from Grades 4–12, this is an opportunity for the candidate to discuss with the examiner a topic of their choice, which has been prepared in advance. Candidates may bring photographs or objects into the examination to illustrate their chosen topic. The free choice of the topic allows candidates to bring into the examination room examples of their own interests and specialisations.
Interactive task	From Grade 7 to Grade 12, candidates are encouraged to take responsibility for leading the interaction with the examiner. The examiner provides an oral prompt and then allows the candidate to initiate and direct the conversation.
Topic presentation	At the Advanced stage (Grades 10–12) candidates make a formal presentation to the examiner, which provides the basis for the ensuing Topic discussion.
Listening task	Also at Advanced stage a discrete listening task is introduced; this is designed to test the candidate's ability to understand the spoken word and make reasoned inferences from a short prompt. It is intended to be solely a test of listening.

It is important to note that though there is a framework to the examination, the examiner does not work to a script (apart from the rubrics to introduce each task). Each examination event is therefore individual. The test-taker has a real interaction with the examiner and has the opportunity to participate in a genuine communicative conversation, albeit within an examination context.

The Integrated Skills in English (ISE) suite of examinations is a relatively new development, designed as a four-skills assessment building on the success and experience of the GESE suite. The oral interview is still central to the examination (forming 50% of the assessment). In addition there are two written

components: a portfolio of work prepared by drawing on a range of resources and media, which is assessed by the oral examiner, and a Controlled Written examination under traditional exam conditions.

As its name suggests, the ISE examinations assess the four skills in an **integrated** context. During the interview, the candidate discusses their written portfolio with the examiner. In the Controlled Written examination, one task involves reading a text or texts and then writing about it. This emphasis on the integrated nature of the language skills reflects real life, where it is rare to use any one skill in isolation.

Similarly, the tasks in both the portfolio and written examination are designed to be similar to the types of tasks which learners of English will have to perform outside the examination context. Emails, letters, reports, reviews and stories are all used — there is no explicit item-level testing of grammar or vocabulary.

No other widely-available English language examination utilises formats like these. The emphasis on realistic tasks, unscripted conversations and the opportunity for the candidate to bring their own interests to the exam ensure that every Trinity examination is a unique example of communicative interaction. Students who are able to communicate in English and who have a firm grasp of the language requirements of the level they are sitting will do well in Trinity exams; students who have memorised lists of words or grammar rules tend to struggle to achieve the real communication required.

A unique history

Trinity's ESOL examinations are inextricably bound up with their unique history. Trinity College London traces its roots back to the first external examinations offered by Trinity College of Music in 1877.

The concept of a graded examination grows out of these early exams in music. While studying an instrument, performers learn, practise and progress in their abilities. Once they have reached the level required for a graded examination they sit the exam. Rightly, they expect to pass. Having passed one grade, the learner continues to practise and progress until they have reached the level of the following grade. There is no set timetable for progression and learners are not encouraged to sit an examination they are unlikely to pass.

From very earliest days, Trinity examined internationally. Trinity music examiners would travel (by ship) to countries such as Australia and India to meet musicians and assess their abilities in their chosen instruments. At no time were local examiners used — Trinity has always believed in a small, standardised panel who could, potentially, examine any candidate anywhere in the world.

Over time, Trinity expanded its range of graded examinations to include elocution and drama. From there, it was a logical progression to include assessment in English for those learning it as a second or other language. During this progression, the key components of the graded examinations in music were retained; the step from grade to grade is small, manageable and achievable. The assessment is based on the performance a student gives on the day of the exam. Students are not encouraged to sit an exam for which they are unprepared. Each examination is designed as a step on a ladder towards full mastery: of an instrument or the English language.

Trinity College London is proud of the long history of examinations in the performance arts which have led to these unique examinations in English language. This background provides a unique perspective on the language and how it is used in everyday situations.

A unique philosophy

Trinity ESOL examinations are the product not only of a long and distinguished history of examining but also of a specific philosophy of learning and assessment. Trinity's examinations are learner-centred, motivational and emphasise the practical aspects of using a language rather than the theoretical aspects of knowing a language.

In every Trinity ESOL examination, the learner is the most important person. Wherever possible, the Trinity examiner will travel to the candidate's place of learning to meet them and conduct the exam. It is the examiner who is a visitor in the exam room, not the candidate. This encourages the candidate to feel 'at home' and comfortable during their examination.

The examiner's role during the examination is to encourage the learner to demonstrate what they can do in English. That is why the Topic phase (discussion and/or presentation, depending on the examination level) is central to every Trinity exam from Grade 4 or ISE 0 upwards. The topic is the candidate's opportunity to present and talk about a subject in which they are genuinely interested and which they

have prepared before the exam. The candidate's interest in their topic area gives them the motivation to learn the vocabulary related to their topic and the enthusiasm to discuss their topic both in the classroom before the exam and with the examiner during the exam itself.

It is also important to Trinity that wherever possible the learner should lead in the examination rather than merely respond to prompts from the examiner. A Trinity oral examination is intended to be a real, interactive conversation — not an interrogation or 'question & answer session'. As well as the Topic phase, this is demonstrated clearly in the Interactive task phase, which is introduced at Grade 7 and ISE II. In the Interactive task, responsibility for leading the conversation is passed from the examiner to the candidate. This role-reversal places the learner at the heart of the examination and gives them the opportunity to direct the course of the conversation and demonstrate their ability to engage someone in a real dialogue at the level of their examination.

Ultimately, what guarantees that the learner is central to every Trinity ESOL examination is the fact that the examination is not scripted. Each exam follows a format which is clearly set out in the syllabus, but every single examination is different. The candidate's input into the conversation will determine the course the interaction takes. Trinity's examiners frequently comment on how much they learn from the candidates they meet. The candidates bring themselves into the examination room, they interact with the examiner and real communication takes place.

A crucial element of the philosophy underlying Trinity's ESOL examinations is that they are designed to have a positive and motivational effect on the candidates. The unique 12-grade system allows learners to receive certification of their abilities from the very earliest days of their English-learning career. This confirmation that you have achieved an identifiable level of English, coupled with the positive experience of the examination itself, encourages learners to continue their learning and to aim for the next step in the Trinity ladder towards proficiency in the language.

With 12 grades of spoken examination, there is an appropriate level for each learner. Students are not encouraged to sit exams which they may fail. Trinity is firm in its belief that it is better — for the learner — to sit an exam which they are likely to pass with Merit or Distinction. This positive reinforcement makes it more likely that they will continue to learn English. Too often, examinations are seen as end results only: the exit point from a course of study or an achievement needed for another purpose (such as university entry). While the benefits of achieving international certification should not be discounted, what this tends to encourage is a feeling that one has 'finished' with English and can forget what has been learned in order to achieve the certification. Trinity's suites of ESOL examinations are designed to avoid this feeling by showing that there is a higher level available and achievable to all learners. In this way, Trinity promotes the lifelong learning of English as a language — rather than the learning of English merely as a means to another end.

The examination itself is also designed to be motivational. Trinity's oral examiners are trained to be friendly and welcoming, and to put the candidate at ease. The examination room should be viewed as an opportunity to interact with a new person and show how much you can do — not a place where your knowledge is tested and you are marked down for everything you do not know. It is an unusual thing to say, but it is absolutely true that candidates leave their Trinity examination having enjoyed the experience. They have seen that the English they have learned in the classroom works in a real-life situation, and they are encouraged to learn more so that the next time they sit a Trinity exam they can do even better.

Similarly, the portfolio component in ISE has the same motivational aim. The candidate has the opportunity to present the best written work they can produce to the examiner. By taking time to research, draft, re-draft and edit their work, candidates develop useful writing skills, become more independent in their learning and develop critical thinking skills. In the oral component of ISE the candidate has the chance to discuss their portfolio with the examiner, covering not only the content but the process of writing it.

The positive motivation does not end in the examination room. Because the oral examiner marks the examination as it takes place, the results of the oral examination are available immediately at the end of the exam session. Students do not have to wait for weeks or months to find out how they performed in the exam. Once the session is completed, Report forms for each candidate are given by the examiner to the Centre Representative. The Report forms indicate how well the candidate performed in each task, suggest areas for improvement for the individual candidate, and give the overall result of the oral

examination. As candidates are encouraged to sit the examination at the correct level for them, pass rates are legitimately high so students enjoy knowing their result immediately and go back to class motivated to continue their learning.

What separates Trinity's examinations from every other widely-available English language exam is the emphasis on the practical ability to use the language, rather than the theoretical ability to demonstrate that you know the language. No Trinity examination hinges on whether a candidate knows a particular word or phrase. Memorising long lists of words and their meanings will not give success in a Trinity examination. The Trinity examinations are assessed according to Task Fulfilment performance descriptors. A language should not be learned by rote; a language should be learned to be used and, in a Trinity examination, the learner uses what they have mastered in the classroom.

A unique impact

It is in the classroom where Trinity's unique impact is most clearly seen. This impact can be summarised as two distinct but related effects.

Firstly, in a classroom preparing for a Trinity examination the teacher is free to teach English as a language to be used in real life situations. The teacher does not have to spend every available moment in exam preparation, teaching skills, vocabulary and techniques which will never be used outside the examination situation. Teachers who help prepare students for Trinity exams enjoy and appreciate the fact that they are able to teach real communicative skills; grammar becomes a tool to assist in accurate communication, not a straitjacket to constrict the teacher and learner.

An English teacher should be able to feel that they are imparting life skills to their students. They should not have to feel that they are preparing students only for an end-of-term examination, which becomes an 'end-of-learning-English' examination. The skills which are necessary to succeed in a Trinity ESOL examination will aid students throughout their English language learning careers — and will encourage them to continue learning English beyond the classroom.

The second major impact of Trinity exams in classroom teaching is that teaching has to become communicative. Learners must become active participants in their learning. They will not succeed in Trinity examinations unless they have practised communicating while they have been learning. Because of this, a Trinity classroom is a noisy classroom!

Teachers find that they have to use a wide variety of approaches and methods in their teaching. They have to encourage their students to speak to each other, to form opinions, to express and defend their opinions — in short: to communicate as effectively and competently as possible.

It is Trinity's firm belief that our unique examinations in English have contributed to an ongoing change in the way the English language is taught around the world. As more and more learners take Trinity examinations, more and more teachers are embracing communicative methods of teaching English. Teachers are talking to their students. Students are talking to their teachers. Students are talking to each other. Trinity College London helps people communicate. We are building on 130 years of experience in assessment and we are encouraging people throughout the world to communicate with each other.

We at Trinity are proud of our examinations and this is why we commissioned the independent research project which is outlined in the following report. The Common European Framework of Reference shares much of the philosophy underpinning Trinity's examinations. The emphasis on communicative ability over linguistic knowledge and the realisation that language learning is a life-long task are common to both Trinity and the CEFR. As the CEFR has become more widely used and understood in classrooms, textbooks and examinations it is important that any person or organisation who makes a claim of calibration with the CEFR is able to substantiate that claim.

Trinity College London's syllabuses for the GESE and ISE examination suites contain claims that each examination level has been mapped to a level of the CEFR. It is important to Trinity that users of our examinations can have confidence in these mappings. That is why we decided to submit our exams to the process described in this report. We were also aware that the unique nature of Trinity's exams would mean that we would be able to contribute to the ongoing work of the Council of Europe to create a Manual for other examination boards to help them accurately calibrate their exams to the CEFR. This report has been submitted to the Council of Europe and the insights generated by this project will be used in the revision of the current version of the Manual for relating language examinations to the CEFR.

Users of Trinity exams — candidates, teachers, schools, universities and employers — can be confident that the mappings in our syllabuses are grounded in the empirical results of this project. They can be confident that holders of Trinity qualifications have demonstrated the ability to communicate in English at the CEFR level specified for their qualification.

I would like to record, on behalf of Trinity College London, my sincere thanks to Spiros Papageorgiou and his research team for the hard work and expertise they contributed to this report.

Clinton Rae
Director of Language Examinations
Trinity College London

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.

Spiros Papageorgiou
Department of Linguistics and English Language
Lancaster University

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). 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 take 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 ongoing 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.).

Table 1. The structure of the GESE suite

Initial	Elementary	Intermediate	Advanced
Grades 1–3	Grades 4–6	Grades 7–9	Grades 10–12
5–7 mins	10 mins	15 mins	25 mins
			Topic presentation Topic discussion
		Topic presentation and discussion	Listening Task
	Topic Discussion	Interactive Task	Interactive Task
Conversation	Conversation	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 2).

Table 2. The structure of the ISE suite

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

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. 78).

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. 81) 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 3,672 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 α (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	
Speaking 1	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 κ (Kappa) calculates exact agreement by also taking into account agreement by chance, which can 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			
	Mean*	Min	Max	Mean	Min	Max	N
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 κ 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 κ, 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 κ 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.

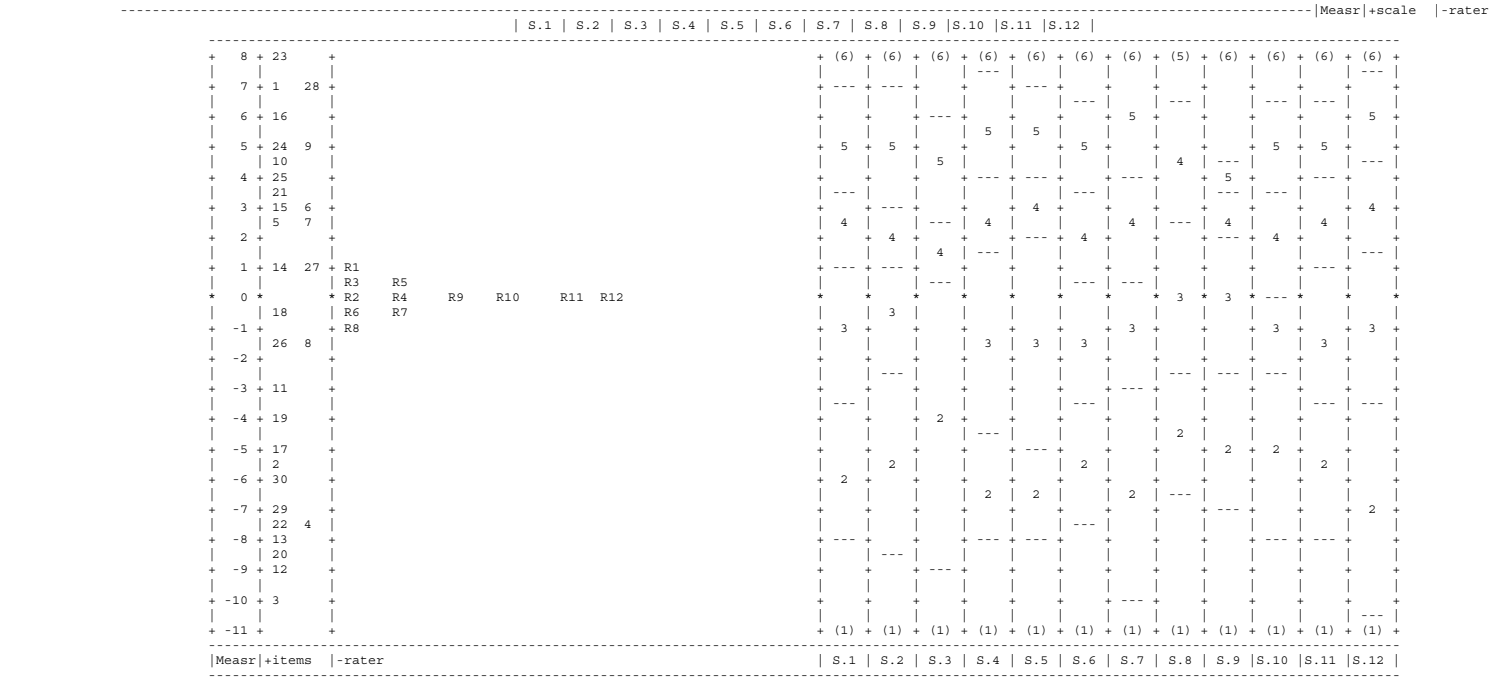
2.2.2 Analysis of judgements using the Rasch model

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 ange	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