TRINITY
COLLEGE LONDON

Speaking and Listening test:

The Graded Examinations in Spoken English

Theoretical background and research

Introduction

The Graded Examinations in Spoken English (GESE) is a suite of oral proficiency tests which assess general English ability at 12 levels. This test provides a valid and reliable scheme of assessment through which learners and teachers can measure progress and development, whether for educational and vocational purposes or as a leisure activity (Trinity College London 2010, p. 5) http://www.trinitycollege.com/site/?id=368. The tests assess communicative competence as defined by Canale and Swain (1980) and Canale (1983). They take the form of direct interviews between an examiner and an individual candidate.

This document has two purposes:

- to present the rationale and theoretical background underpinning the development of the the GESE and
- to describe the research that Trinity College ('Trinity') has conducted in the past three years, to investigate the validity of this test.

The document is divided into four sections. The first section provides the historical background to the GESE. The second section describes the GESE examination tasks and the rationale behind them. The third section presents a discussion of the theoretical underpinnings of the GESE, and because it is based on the GESE. The fourth section summarises the validation studies carried out by Trinity in the past three years to validate both these tests.

Section 1 - Historical background to the GESE

1.1 Overview

When the GESE was developed in 1938 it resembled an elocution test rather than a meaningful exchange between examiner and candidate. The examination evolved gradually and by the 1980s it reflected the influence of more communicative teaching methodology. The grading system matched the system employed in Trinity's examinations in music and other performing and creative arts, which was devised to measure individual progress in incremental steps, through a criterion-referenced rather than norm-referenced system of assessment.

The GESE examination consists of four developmental stages (Initial, Elementary, Intermediate and Advanced), with three grades at each stage, making 12 grades in all. A detailed description of the stages and grades can be found in Table 1 and Section 3.

Table 1. The Graded Examinations in Spoken English (GESE)

Stage	Grades	CEFR levels	Timing (minutes)	Conversation	Topic Discussion	Interactive Task	Topic Presentation	Monologue Listening
Initial	1-3	Pre-A1 -A2	5-7	✓				
Elementary ISE 0 and I	4-6	A2 – B1	10	✓	✓			
Intermediate ISE II	7-9	B2	15	✓	✓	✓		
Advanced ISE III and IV	10-12	C1-C2	25	✓	✓	✓	✓	✓

This fine tuning of assessment level reflected, and still reflects, a belief that evidence of successful learning is an important motivator for future progress. Allowing the individual to choose the level and timing of an assessment and conducting the assessment in a supportive environment provides candidates with the maximum opportunity for success – what Swain (1984) called 'bias for best'.

No formal construct was specified when GESE first appeared; however, the specifications, in the form of a syllabus, have always been available to potential candidates and other users.

1.2 Key developments 1960 – 2010

The most significant developments in GESE are summarised in Table 2, and are explained in more detail in the paragraphs that follow.

Table 2: Significant developments in GESE

Time Period	**Nature of development**
1960s – 1980s	Move from grammar-based test towards communicative competence.
1980s - 2000	Text Discussion is replaced by a Topic Discussion, to avoid memorisation and encourage more natural and spontaneous communication.
2000 – 2004	Introduction of an Interactive Task to assess functional language and the candidates' ability to initiate and maintain conversation. Introduction of the Listening task to assess monologue listening. Revision of the assessment criteria
2005 – 2007	CEFR calibration project. Specifications review to align with CEFR scales.

In the 1960s the GESE resembled an oral grammar knowledge test. A typical item might be:

Examiner: Give me a sentence using *unless.*

Candidate: I'll go to the party *unless* I have to work.

The suite remained heavily grammar-based until the 1980s, when changes were introduced to reflect current trends in communicative language teaching. From Elementary stage upwards candidates were asked to talk about a book or article they had read (Text Discussion) rather than produce grammatically accurate sentences. The Text Discussion task was eventually replaced, however, as some candidates relied on memorisation rather than displaying spontaneous language.

Further substantial revisions were made between 1980 and 2004. The Text Discussion task was replaced by a Topic Discussion task, in which candidates could select and prepare a topic of their choice to discuss with the examiner. The revised task was intended to encourage more natural and spontaneous interaction between candidate and examiner, by being more 'personalised' and allowing the candidates to express their own views and opinions. The Interactive task was introduced at the Intermediate and Advanced stages. The focus of this task was on active listening and communication skills. The task required the candidate to take the initiative by asking questions and maintaining the interaction with the examiner. The language focus was on functional language rather than, for example, accurate grammar production.

In this period the assessment criteria were also modified to focus on the more holistic criterion of task fulfilment. Task fulfilment was defined as the candidate's use of the language listed in the specifications, i.e., the communication skills, functional language, grammar, lexis and phonology, for each grade being examined. The decision to alter the assessment criteria reflected the communicative approach of the GESE.

The GESE suite was calibrated to the CEFR during a two-year period between 2005 and 2007. A full account of the calibration process can be found in Papageorgiou (2007), available on the Trinity College London website. The resulting calibration can be seen in Table 3 below. Some changes were made to the language criteria specified for particular GESE grades; for example, candidates now ask the examiner a question at Grade 2 (CEFR A1).

Table 3. The Common European Framework of Reference and The Graded Examinations in Spoken English and The Integrated Skills in English examinations

Common European Framework of Reference (CEFR)	**Graded Examinations in Spoken English (GESE)**
-	Grade 1
A1	Grade 2
A2	Grade 3 Grade 4
B1	Grade 5 Grade 6
B2	Grade 7 Grade 8 Grade 9
C1	Grade 10 Grade 11
C2	Grade 12

As mentioned earlier, the GESE specifications are reviewed every three years or so in order to reflect developments in language teaching and the needs of the test takers. The most recent revision was in 2010. Qualitative feedback from approximately 300 examiners was collected, collated and analysed. A panel of senior examiners plus Trinity staff from the ESOL test development and research teams determined the changes to be made. In this revision some of the subject areas for conversation were altered to be more appropriate for the candidature. It was also decided to revise the way information about the examination tasks was presented in the specifications, to make it more accessible to teachers, candidates and examiners.

1.3 Current and future developments

The GESE specifications are due for its next revision in 2016. The changes will be informed by recent and current research on the GESE construct, the Interactive Task, the Listening Task, interactive listening comprehension, holistic rating scales and examiner behaviour. Reports from Trinity's pre- and post-testing activities will also contribute to the content revisions. Further changes may also be made in the way we present information in the Exam Information Booklet, in order to address the specific needs of particular stakeholder groups.

Section 2 - The GESE examination

This section gives a brief overview of the current GESE. Further details can be found in the GESE Exam Information Booklet (http://www.trinitycollege.com/site/?id=368)

2.1 Description of the GESE Suite

The graded examinations are designed for speakers of languages other than English. The graded system sets realistic objectives in listening to and speaking with English speakers. The 12 grades provide a continuous measure of linguistic competence and take the learner from absolute beginner (Grade 1, below A1 CEFR) to full mastery (Grade 12, C2 CEFR). The grades are organised in four development stages (see Table 1 above). The language and skills assessed for each grade are common components in most English language curricula and they do not require specific examination preparation textbooks. Trinity provides teachers with free guides to exam content and preparation on the GESE pages. Examples can be found at: http://www.trinitycollege.com/site/?id=368

GESE is a face-to-face oral interview between a single examiner and a single candidate. The examination simulates real-life exchanges in which the candidate and the examiner pass on information, share ideas and opinions, and debate topical issues. Examiners are not provided with a prescribed script for the interview, but rather base their questions on a test plan which they have developed themselves using the language of the grade. This helps examiners to elicit the language and communication skills required by the grade the candidate is sitting for, while allowing them to respond to the candidate's contributions.

2.2 Rationale and description of the GESE test tasks

As candidates progress through the grades they are required to demonstrate a greater range of communication skills and more complex language functions, grammar structures, lexis and phonology. A new task is introduced at each stage to assess these competencies. The aim behind all the tasks in the GESE suite is to elicit natural interaction between the candidate and examiner in order to give as accurate a reflection as possible of how the candidate might perform in the real world. The GESE tasks require candidates to use interactive listening skills in all but the Advanced Listening task. Interactive listening skills are needed to respond appropriately to the examiner's contributions in the discussions.

A brief description of the tasks is given in Table 4, and more detailed discussion follows.

Table 4. Description of GESE tasks

GESE tasks	Time	Prepared beforehand	Spoken Interaction demands	Interactive Listening demands
Topic Discussion Task	5 mins	✓	demanding	less demanding
Topic Presentation (Advanced Stage only)	5 mins	✓	no interaction	N/A
Interactive Task	4 – 5 mins	spontaneous	demanding	demanding
Listening Task (Advanced Stage only)	3 mins	spontaneous	less demanding - candidate supplies short answers.	demanding
Conversation Task (inc. ISE Portfolio discussion)	5 – 7 mins	spontaneous	demanding	demanding

2.2.1 The Topic Discussion

Rationale

The rationale for the Topic Discussion task is to generate a natural exchange of ideas and opinions between candidate and examiner by means of an information gap. Candidates have complete autonomy when preparing for this section and can choose any subject they wish to discuss with the examiner. Allowing candidates to prepare their topic in advance gives them the opportunity to demonstrate what they can do with English when they are given time to anticipate typical questions and to use resources such as interactive listening skills to answer the questions apprpopriately.

Task Description

The candidate completes a Topic 'mind map' (http://www.trinitycollege.co.uk/site/?id=1980), which the examiner will use as a basis for the discussion. Candidates may also create materials to illustrate their topic. The Topic mind map and materials are not assessed. The Topic Discussion task provides the candidate with the opportunity to show they can link sentences together to talk about a subject at some length. This task matches the CEFR Coherence and Cohesion descriptor:

- A2: …*can link simple sentences in order to tell a story or describe something*
- B1: *Can link a series of shorter, discrete simple elements into a connected, linear sequence of events*
- B2 *Can use a limited number of cohesive devices to link... utterances into clear, coherent discourse.* (CEFR 2001, p.125).

It also corresponds to the CEFR B1 Informal Discussion descriptor: *Can give or seek personal views and opinions in discussing topics of interest*. At B2 *Can express his/her ideas with precision*… (CEFR 2001, p.77)

The Topic Discussion task changes at the Advanced Stage, when the candidate gives an uninterrupted Topic Presentation before the candidate opens the discussion.

2.2.2 The Topic Presentation (Advanced Stage only)

Rationale

An oral presentation task was selected since it is a common real-world task - for example, in job interviews. The task gives candidates the opportunity to display their command of the language in an uninterrupted and formal situation. Candidates are expected to present abstract concepts clearly and concisely over a series of connected long turns. It is assumed that candidates at this level (CEFR C1-C2) will be well-motivated and have particular reasons for wanting to be fluent in English. Normally, candidates will be mature and experienced enough to handle abstract concepts and to contribute to discussions on matters of major importance in today's world.

Task Description

Candidates give an uninterrupted 5-minute formal presentation on a subject they have chosen themselves and prepared beforehand. This task corresponds to the CEFR C1 and C2 Thematic Development descriptor: *Can give elaborate descriptions and narratives, integrating sub-themes, developing particular points and rounding off with an appropriate conclusion* (CEFR, p.125). It also matches the C1 Addressing Audiences descriptor: *Can give a clear well-structured presentation of a complex subject, expanding and supporting points of view at some length with subsidiary points, reasons and relevant examples* (CEFR, p.60).

2.2.3 The Interactive Task

Rationale

The rationale for the Interactive task is to produce an authentic exchange of information and opinions. The candidate is expected to demonstrate control over not only the use of the language functions (Intermediate Stage) but also language form (Advanced stage), in an integrated and meaningful way. Interactive listening skills are required to enable the candidate to respond appropriately and to direct the conversation. The prompt, provided by Trinity, is not known by the candidate beforehand and is designed to elicit sociolinguistic and strategic skills as well as linguistic skills. The examiner does not work from a prepared script, but creates a test plan with a backstory for each prompt which will (ideally) provoke certain kinds of questioning and comments from the candidate. The GESE Construct Project (O'Sullivan, 2010) showed that candidates did indeed exhibit sociolinguistic and strategic competencies. See Section 4 of this report for more details.

Task Description

This task is introduced at the Intermediate Stage, Grade 7 (CEFR B2). The Interactive task begins with a prompt from the examiner, which describes specific situations or issues which the candidate should try to respond to by asking questions of the examiner, and expressing their own opinions and ideas. The task requires the candidate to initiate 'turns' in the conversation and manage the direction of the interaction. The Interactive task corresponds to the CEFR B2 Taking the floor (turntaking) descriptor: *Can initiate, maintain and end discourse appropriately with effective turntaking*. (CEFR, p.86). It also matches the CEFR B2 Understanding a Native Speaker Interlocutor descriptor: *Can understand in detail what is said to him/her in the standard spoken language.* (CEFR, p.75)

A set of prompts is provided by Trinity for each grade annually. All the prompts are pretested on an appropriate cohort and approved by an international bias review board (see: Test Production and Pretesting documents, Final Report, Appendices 2 and 3).

2.2.4 The Listening Task (Advanced Stage only)

Rationale

All the GESE tasks, apart from the Advanced Stage Topic Presentation, require the candidate to use interactive listening skills. Interactive Listening Skills are defined as the listening skills which support effective interaction (Ducasse and Brown 2009). The rationale behind the discrete Listening Task is to assess monologue listening skills using a pre-constructed non-stop monologue – the kind of input that is common in real-world tasks such as listening to

lectures or to the radio. It is expected that the candidates at this level are almost fully proficient, so they should be able to react to any spoken language in any context.

Task Description

The Listening Task is introduced at the Advanced Stage. The input is non-specialist and does not relate to the specific subject areas provided for the Conversation task. The candidate is required to give only very brief verbal responses indicating comprehension rather than productive ability. The candidate needs to show recognition of the context, participants and register.

The examiner reads three short spoken passages to activate the use of high-level listening skills, such as deduction, prediction and inference. This activity corresponds to:

- CEFR Global Descriptor for C1: *Can understand a wide range of texts, and recognise implicit meaning*
- C2: *Can understand with ease virtually everything heard or read* (CEFR, p.24).
- CEFR C2 Overall Listening Comprehension descriptor: *Has no difficulty in understanding any kind of spoken language, delivered at fast native speed* (CEFR, p.66).

2.2.5 The Conversation

Rationale

The rationale for the Conversation task is to reflect a realistic exchange of information, ideas and opinions. The subject areas have been carefully selected to offer a progression from the familiar to the less familiar and from the 'concrete' to the 'abstract', and to elicit more sociolinguistic and strategic skills at each successive grade.

Task Description

The Conversation task is common to all GESE grades. It is an informal discussion on two of the Subject Areas listed in the specifications for each grade. At Initial Stage (Grade 1 – Grade 3) the focus is on the exchange of basic personal information and familiar matters. This corresponds to CEFR Overall Spoken Interaction descriptor A1*: Can ask and answer simple questions, initiate and respond to simple statements on very familiar topics.* (CEFR, p.74).

At each grade the candidate is expected to take more responsibility for initiating and maintaining the conversation. This development of conversation skills corresponds to the CEFR Conversation descriptor:

- A2: *Can participate in short conversations…on topics of interest.*

- B1: *Can enter unprepared into conversation on familiar topics. Can maintain a conversation or discussion* (CEFR, p.76).
- B2: *Can engage in extended conversation on most general topics.*

From Grade 4 to Grade 11 the examiner selects two Subject Areas from the list published in the Exam Information Booklet, specifications section. At Grade 12 there is no list since C2 level candidates are expected to be able to discuss a wide range of topics of general or topical interest. This matches the C1 and C2 Informal Discussion descriptor: *Can easily follow and contribute to discussion... even on abstract, complex and unfamiliar topics* (CEFR, p.77)

2.3 The assessment criteria

The candidate's performance in the examination is measured by means of one overall criterion,Task Fulfilment. Examiners use this criterion as they judge each task in the examination. In the GESE Task Fulfilment includes the following factors:

- competence in the communicative skills listed
- coverage of the language functions listed
- coverage of the grammatical, lexical and phonological items listed
- accuracy in the use of the grammatical, lexical and phonological items listed
- appropriacy of the grammatical, lexical and phonological items used
- fluency and promptness of response appropriate for the grade.

Detailed Performance Descriptors for Task fulfilment are available at http://www.trinitycollege.com/site/?id=3112

The examiner assesses the candidate's performance in each task of the examination by awarding a letter grade A, B, C or D. In simple terms, these levels can be interpreted as follows:

- A — Distinction an (excellent performance)
- B — Merit (a good performance)
- C — Pass (a satisfactory performance)
- D — Fail (an unsatisfactory performance)

Section 3 - Theoretical background

As described in Section 1, GESE has undergone a number of revisions over the years, responding to changing trends in English language teaching and testing. Section 3 summarises the theory underlying the GESE (its construct) and Section 4 presents the research Trinity has recently carried out in order to enhance its construct validity.

3.1 Literature review

The GESE is a performance examination which reflects real-life interactions, with the focus on communicative competence and meaning. The theories underpinning the examination were being explored and articulated from the 1970s onwards. Topical theories were the integration of skills to reflect real-world usage and the building of meaning between the participants. Two points emerged from the discussions in those years: the importance of interaction, and an emphasis on language use for communication.

The importance of interaction was argued by Savignon, who claimed that communication was 'Dynamic rather than…static… [depending on] the negotiation of meaning between two or more persons' (1972, as cited in Canale and Swain,1983, pp.8-9). Some years later Kramsch (1986) considered communicative interaction to include 'anticipating the listener's response and possible misunderstandings, clarifying one's own and the other's intentions and arriving at the closest possible match between intended perceived and anticipated meanings.' (1986, p.367).

The second point to emerge was articulated by Widdowson (1978), who claimed that 'in normal conversation speakers will focus more on language use than grammar' (cited in Canale and Swain 1980, p.5). Van Ek (1976) was also concerned with speakers' abilities to communicate. He proposed a Threshold Level for linguistic competence – the point where learners could survive independently in the target language (cited in Canale and Swain 1980, p.9).

Canale and Swain proposed a model of communicative competence in 1980, which Canale expanded in 1983. The expanded model divided communicative competence into grammatical, sociolinguistic, discourse and strategic components. Brief explanations for each type of competence can be found in Table 5 below.

Table 5 Theoretical bases of communicative approaches to second language teaching and testing. Canale & Swain (1980), and Canale (1983)

Component	Definition
Grammatical	Knowledge of lexical items and of rules of morphology, syntax, sentence-grammar semantics, and phonology
Sociolinguistic	Knowledge of the socio-cultural rules of language and of discourse
Discourse	Ability to connect sentences in stretches of discourse and to form a meaningful whole out of a series of utterances
Strategic	The verbal and non-verbal communication strategies that may be called into action to compensate for breakdowns in communication due to performance variables or due to insufficient competence

Bachman (1990) and Bachman and Palmer (1996) proposed a model of communicative language ability (CLA), which posited two major competences: language competence and strategic competence (1990, p.85). In contrast to Canale and Swain, Bachman and Palmer envisaged strategic competence as the mobilisation of higher-order metacognitive strategies, such as assessment and planning (1990, p.102, and Bachman & Palmer, 2010, p.48).

An aspect of communicative language testing which was not present in language tests in the 1970s and 1980s was authenticity. Bachman and Palmer (1996), building on Bachman (1990), proposed two types of authenticity. They used the term 'authenticity' to refer to 'the degree of correspondence of the characteristics of a given language test task to the features of a TLU (Target Language Use) situation' (p.23). They saw this notion as crucial to task design 'because it relates the test task to the domain of generalization to which we want our score interpretations to generalise' (p.23). They used the term 'interactiveness' to refer to 'the extent and type of involvement of the test-taker's individual characteristics in accomplishing a test task' (p.25).

Swain (1985) considered that the content of communicative language tests should be motivating, substantive, integrated and interactive. She recommended using opinions and/or controversial ideas, plus new information in order to create a natural information gap to stimulate interaction (as cited in Bachman 1996, p.320). Norris, Brown, Hudson and Yoshioka (1998) developed the argument further by stating that performance tests should be as authentic as possible.

The challenge, however, is not just to state the theory underlying the GESE suite but also to investigate the validity of the claims made that the tests are assessing relevant aspects of the theory.

Section 4 - Validation studies

Trinity has commissioned a number of studies over the past three years to investigate the validity of the inferences we claim can be made through our examinations. In this section we will describe five of the most important studies, and we will then discuss how Trinity's now routine procedures such as pretesting, monitoring and post-test analysis are contributing to our validation efforts.

4.1 Key Validation Studies

4.1.1 GESE Construct Project

Professor Barry O'Sullivan (Roehampton University) was invited to design a project which would enable Trinity to investigate whether the GESE examination tapped the four competences described by Canale and Swain, and if so, in what measure. We decided to apply a framework derived from Weir (2005, and O'Sullivan & Weir, 2010) to analyse the GESE specifications, the performance descriptors in our rating scale, and test-taker performance at all levels of the examination.

The analysis was carried out by Professor O'Sullivan, two other external consultants specialising in second language acquisition and pedagogy, and two testing specialists from Trinity. We felt it important to bring in external consultants so that we could combine the objective observations made by 'fresh eyes' with our own understanding of the intentions behind the examination and of the context we were working in.

The project team analysed three test-taker performances at each of the 12 levels of the examination, which totalled 36 performances in all. They analysed each of the components of the examination at all the levels they appeared: the Conversation task at Grades 1 - 12, the Topic Discussion at Grades 4 - 12, the Interactive Task at Grades 7 -12, and the Topic Presentation and the Listening Task at Grades 10 - 12.

The details of the framework we used are too complex to be described in this short summary (see Weir 2005, and O'Sullivan & Weir 2010 for further explanation), but the main foci were the test-taker (individual and cognitive characteristics), the test system (test tasks and administration), and the scoring system (theoretical fit, accuracy of decisions and value of decisions).

The general conclusion of the project was that the GESE was successfully tapping all four of the Canale and Swain competences. Grammatical competence is displayed from Grade 1 (CEFR pre-A1), when candidates are asked to use short answers to simple requests for information. Discourse and strategic competences emerge from Grade 3 (CEFR A2), and are clearly displayed by the end of the Elementary Stage, when candidates have to give

information in a series of sustained turns in the Topic Discussion. At this stage candidates start to play a limited part in initiating and maintaining the interaction in the Conversation Task. Sociolinguistic competence begins to appear from Grade 4 (CEFR A2.2), and is clearly displayed at Intermediate Stage, when candidates are expected to engage the examiner in discussion and maintain the interaction. By the Advanced Stage all of the competencies are fully developed and evident throughout the test-taker's performance.

As has been discussed above two unique features of Trinity examinations are their unscripted nature and their focus on interaction. Two recent studies have looked into the demands these features places on candidates.

4.1.2 GESE Grade 7 - Interactive Listening

Nakatsuhara and Field (2012) analysed the interactive listening skills placed on candidates in all tasks at Grade 7. Interactive listening skills were defined as the skills which support effective interaction (Ducasse and Brown, 2009). As mentioned earlier, The GESE is highly structured but not scripted and is meant to elicit natural and spontaneous interaction. This study investigated the nature of the skills candidates require to cope with this kind of activity. The research drew on Weir's (2005) *Socio-cognitive framework for validating language tests*, which was also utilised by O'Sullivan (2010) for the GESE construct project. The data consisted of recordings of 20 examiners, each interacting with a low-level candidate and a high-level candidate. The examiners' interventions were analysed for lexical complexity, syntactic complexity, informational density, number and mean length of interventions and purpose of interventions. The analysis had a dual function of acting as a validity check on the examiners' utterances. The results showed that each of the three Grade 7 tasks provided the candidates with different listening challenges and tapped into the candidate's interactive listening skills. There was clearly a connection between the speaking skills demonstrated by the candidates and their abilities as listeners to respond to turns of the examiner. This outcome validates our claims that each of the tasks adds a different dimension to the listening ability we are assessing.

It was also clear that the different tasks impose different demands on the test-taker's listening abilities. The Topic Discussion is less demanding than the other two tasks since the test-taker prepares in advance and examiner questions can be anticipated. The Interactive and Conversation Tasks are more demanding, particularly when the examiner changes sub-topic. This is most notable in the Conversation Task, which is directed more by the examiner than the Interactive Task.

4.1.3 GESE Interactive Task – Pragmatic Demands

Our communicative approach to testing inevitably requires some pragmatic competence . This is particularly evident in the Interactive Task where the candidate is required to elicit and manage the examiner's sequential revelation of important details about a situation. This study investigated how far the test tasks tapped into the functional use of linguistic resources.

Hill (2012) looked at the pragmatic demands placed on candidates in the Interactive Task at Grades 7 to 9. Audio recordings from live examinations were analysed, using a pragmatic framework derived from a range of theorists: (e.g., Austin, 1975; Grice, 1975; Hymes, 1962; Pomerantz, 1984 and Thomas, 1995). The approach was very detailed and analysed each utterance in terms of its meaning on six levels: pragmatic force, discoursal intent, interpersonal intent, the contextual constraints of the interaction, the perlocutionary effect on the hearer in relation to all of the above, and the hearer's response. The process of analysis was iterative, with each utterance analysed for responsiveness, negotiation and initiation in relation to the creation of meaning. Significant points in the interaction were then selected for analysis in relation to the professional management of the encounter.

The research showed evidence of clear pragmatic foundations in the Interactive construct and the design of the prompts. The Interactive task specifically tests and assesses a candidate's ability to correctly infer implied meanings. This was a short study and a complex issue such as pragmatics requires more detailed analysis, but nevertheless the findings supported the validity of our claims about the tasks in question.

4.1.4 GESE Listening Task

Research has also been undertaken to study the demands placed on candidates in the Listening Task in the Advanced stage. Brunfaut and Révész (2009) investigated the effects of a group of task factors on advanced ESL learners' actual and perceived listening performance. They examined whether the speed, linguistic complexity, and explicitness of the GESE Grade 10 Listening Task, Type 1, influenced comprehension. The data consisted of responses from 68 students to 18 listening text prompts for Grade 10. A post-test perception questionnaire was completed and nine students took part in a stimulated recall exercise.

Rasch and regression analysis were used to investigate task difficulty and its relationship to text characteristics - for example, lexical complexity. Findings indicated that lexical complexity was the key predictor of task difficulty and that syntactic complexity did not have a significant impact on learner performance. Findings from the stimulated recall reported trends similar to those reported for task difficulty. However, it was noted that many of the listening task texts bore more similarity to written rather than spoken texts.

This issue is now being reinforced in Item Writer training and Listening Task Item Writer guidelines.

4.1.5 Young Learners

Since many of the candidates taking the lower grades are very young, we have begun investigating the testing of young learners. The first step in this research was to commission a literature review of foreign language testing of speaking and listening for 5 – 11 year olds (Nikolov 2012). The reviewer was asked to comment on the GESE specifications and video performances of young learners in the light of this research. The recommendations are wide ranging and will be considered should Trinity decide to produce a test specifically for young learners.

The above paragraphs have described special reports commissioned by Trinity to investigate the validity of its examinations. It is important to add that we also gather information as part of our routine validation checks.

4.2 Impact studies

Trinity regularly conducts impact studies on its Speaking and Listening examinations. The most recent was carried out in July 2012 in Italy. Headteachers, teachers and students (middle and higher schools) participated. The survey asked about the examination experience, perceptions of the exam (including parents'), and motivational impact. Specific questions for teachers focused on changes, if any, in their teaching methods, class assessment and benefits for their CPD (continuing Professional Development). The results of the latest survey will be published on the Trinity website at the end of September 2012. Previous studies consistently show that students find the test motivating and teachers report an increase in more communicative teaching methodology.

4.3 Quality assurance

As The GESE is an unscripted speaking test it is important to check that Interactive and Listening tasks elicit the intended communicative skills, functions, grammar, vocabulary and phonology listed for each grade. The Trinity Pretesting programme checks that the prompts meet the specifications and gives information on examiner performance. We have a strong monitoring programme, details are provided in Section 4.2.2

4.3.1 Pretesting

The Interactive and Listening tasks pretests are all audio-recorded and both quantitative and qualitative analysis is carried out at item level. Item Response Theory (IRT) is used for the Interactive task to investigate the quality of the items, candidate ability, inter- and intra-rater reliability. Classical analysis is used for the Listening tasks to determine the facility value and discrimination index of the items.

Interactive Task: qualitative validity checks

Interactive task recordings are sent to a monitor who is a Speaking and Listening asssessment expert.[3] The monitor has two duties when listening to performances on the Interactive Task. The first is to check whether the prompts elicit the functions specified in the specifications for the grade being tested. The monitor works with a checklist which includes functions for that specific grade, and for the grades just above and just below. For example, a monitor checking performances on a Grade 8 task will check the functions from Grades 7 and 9 as well.

The second duty is to check whether the examiner's contributions and backstories help to elicit the language and communication skills of the grade. It will be recalled (Section 2.2.3) that the examiners develop their own backstory for each prompt. The Pretest Review meeting reviews all the monitor's (monitors') comments and decides which prompts elicit the required language functions and which backstories were most effective. Key points from the successful backstories are incorporated into the Examiner Training Programme, to help other examiners plan their backstories.

Listening task

The listening task requires a single-word response or a very short answer. In order to pretest items as efficiently as possible an appropriate sample of candidates completes all 30 listening items for a particular grade at one sitting. The prompts are also tested on candidates above and below the grade level of the prompt - for example a Grade 11 candidate will answer Grade 10 and Grade 12 prompts. The responses are statistically analysed using Classical Analysis. In the Pretest Review meeting the prompts which are too easy, difficult and/or do not discriminate properly are discarded.

[3] For further explanation of how Trinity considers an 'expert' please see Appendix 3, Pretest Processes Overview document, p.24

The remaining prompts and answers are scrutinised to check that the prompts elicited the predicted response. Those which do not are returned to the editing process or discarded. Successful tasks are submitted to the Bias Review panel.

4.3.2 Examiner monitoring

Audio-monitoring

All the GESE examinations are audio recorded and 20 – 25% of the examiners are monitored annually. GESE monitors listen to a sample of the examiner's recordings, checking the examiner's language and contributions to the exchanges to see whether the examiner is adhering to the language of the grade. The findings from all the monitors are collated and incorporated into examiner training sessions and the Item Writing process.

Live monitoring

Trinity also has a live-monitoring programme to check that examiners are performing their role satisfactorily. 30% of the examiner panel are live-monitored each year. The monitor spends half a day to a day with the examiner, observing the interaction with candidates and recording their own judgements about the marks that should be given. A feedback session is held at the end of the day, where the monitor and examiner discuss examination techniques, test plans, materials and scores awarded. In total 50% of the examiners are checked every year (20% in audio-monitoring, and 30% in live monitoring), to ensure that the examiners' performances are in line with the GESE construct and that the examiners are following Trinity procedures.

Conclusion

This concludes the description of the GESE test, the validation studies which have been carried out to investigate the construct, and activities which are carried out as part of the Test Production process to check whether the construct is being adhered to. Trinity now has a designated Research and Development department to conduct a full programme of research across all Trinity's examinations and qualifications. Trinity has always recorded a percentage of its examinations but since 2011 all the Trinity Speaking and Listening tests have been digitally recorded. This rich and signficant supply of data will be used in future validity and research studies. We will be carrying out more research and validation studies for the forthcoming GESE review.

References

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

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford, UK: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice.* Oxford, UK: Oxford University Press.

Brunfaut, T., & Révész, A. (2009). Tasks and assessing L2 listening comprehension. Report commissioned by Trinity College London

Canale, M. (1983). On some dimensions of language proficiency. In J.Oller (ED.), *Issues in language testing research,* 33-42.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second lanaguage teaching and testing. *Applied Linguistics, 1* (1), 1-47.

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

Ducasse, A. M. and Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing* 26(3), 423-443

Hill, C. M. (2012). An interactional pragmatic analysis of the GESE interactive task. Internal report for Trinity College London.

Fox, J. (2009). Biasing for best in language testing and learning: an interview with Merrill Swain. *Language Assessment Quarterly, 1*(4), 235-251.

Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal, 70*, (4), 366-372.

Norris, J.M., Brown, J.D., Hudson, T., and Yoshioka, J. (1998). *Designing second language performance assessments.* Honolulu, USA: University of Haiwaii Press.

Nikolov, M. (2012). Testing young learners in a foreign language. Internal report for Trinity College London.

O'Sullivan, B. (2010). GESE construct project. Report on the outcomes of the construc definition procedure. Internal report for Trinity College London.

O'Sullivan, B., and Weir, C. J. (2010). Language testing = validation. In O'Sullivan, B. (ed.) *Language Testing: Theories and Practices.* Oxford: Palgrave.

Papageorgiou, S. (2007). Relating the Trinity College London GESE and ISE examinations to the Common European Framework of Reference. Final Project Report, February 2007 http://www.trinitycollege.co.uk/site/?id=1245

Trinity College London Graded Examinations in Spoken English (GESE) Exam Information Booklet http://www.trinitycollege.co.uk/site/?id=1976

Weir, C. J. (2005) *Language testing and validation: An evidence-based approach.* London: Palgrave Macmillan