HellenicAmericanUniversity

BCCE™

Basic Communication Certificate in English

Standard Setting of the Basic Communication Certificate in English (BCCE™) Examination: Setting a Common European Framework of Reference (CEFR) B1 Cut Score

Technical Report

Standard Setting of the Basic Communication Certificate in English (BCCE™) Examination: Setting a Common European Framework of Reference (CEFR)

B1 Cut Score

Technical Report

Charalambos Kollias

Office for Language Assessment

Hellenic American University

Office for Language Assessment
505 Amherst Street, Nashua, New Hampshire,
NH 03063, USA
T: +1 603-577-8700
e-mail: info@hauniv.edu

Table of Contents

List of Tables

List of Figures

Acknowledgments

I would like to warmly thank Dr. Spiros Papageorgiou, Dr. Richard Tannenbaum, and Dr. Sauli Takala for sharing their professional standard setting expertise with me and answering my queries concerning the CEFR Manual (Council of Europe, 2009) and standard setting procedures, to Gareth McCray for providing support with the rater agreement index (RAI), to Eri Naska for proofing the final version of this report, to Anna Razou for designing the cover page and laying out this report, and to the panelists for participating in the study.

Further thanks go to Dr. Sauli Takala for providing assistance from the onset of the study, providing feedback on an earlier version of this report, and agreeing to be the external reviewer of this report.

Special thanks go to the Hellenic American Union Center for Examinations and Certifications for organizing the meeting and to the Hellenic American University for offering facilities and services to conduct it.

1. Introduction

This report presents the results of a four-day standard-setting workshop organized to set a B1 proficiency level of the Common European Framework of Reference (CEFR) cut score for the Basic Communication Certificate in English (BCCE™) examination. The methodology and the results of this workshop are discussed.

1.1 Standard Setting

Standard setting is a decision making process (Kaftandjieva, 2004) of establishing a cut score (Cizek, Bunch, & Koons, 2004) or "a point on test scale which is used to separate candidates into two categories, each reflecting a different level of proficiency relative to the competency measured under the test under consideration"(Hambleton & Eignor, 1978: 5).

In order for a cut score to be determined, a group of experts (judges) are recruited to take part in a standard setting workshop and to recommend a cut score for a certain examination. The policy committee reviews the documentation of the standard setting workshop and the experts' recommended score and makes the final decision on the cut score.

1.2 The Basic Communication Certificate in English (BCCE™) Examination

The Basic Communication Certificate in English (BCCE™) is a standardized examination for candidates who wish certification at an intermediate proficiency level in English. The examination has been mapped onto the Common European Framework of Reference for Languages (CEFR) as reflecting the content and difficulty of B1 level. The BCCE™ examination tests communicative competency in all four language skills: reading, writing, listening, and speaking. It also contains sections that specifically test grammar and vocabulary resources. The examination focuses on all four domains specified in the CEFR: personal, educational, occupational, and public.

The BCCE™ examination consists of four sections. Section 1: Listening contains 30 multiple-choice questions, with three answer options per item. There are four parts in the Listening section: short conversations with answer choices shown as pictures, short-recorded messages or announcements with one question, longer dialogues broken into three segments with two questions per segment, and a short talk followed by five questions. Section 2: Grammar, Vocabulary, and Reading (GVR) contains 75 multiple-choice items with four answer options per item. There are 25 grammar items, 25 vocabulary items, and 25 reading comprehension items. Section 3: Writing contains two short semi-structured e-mails, 80 to 100 words each. Section 4: Speaking contains a structured oral interaction with an oral examiner. The interaction also involves a visual prompt and a role-play activity.

The BCCE™ examination is developed and scored by the Hellenic American University in accordance with the rigorous international standards of educational measurement. All parts of the examination are written following specific guidelines and test specifications. Items are pre-tested to ensure standardization and psychometric quality. Care is taken that the tests are fair and accessible to all examinees. The BCCE™ examination is administered by local institutions at test centers around the world. The Office for Language Assessment and Test Development (OLATD) at the Hellenic American University works closely with local test centers to ensure secure and reliable administration of all examinations, whenever and wherever they are administered. OLATD provides equal opportunities to all candidates without any discrimination. It is committed to providing equal opportunities for all members of its public and prohibits discrimination

on the basis of race, color, gender, sexual orientation, age, religion, national origin, physical disability, or veteran status.

1.3 Purpose for Setting CEFR Level Cut Scores

The BCCE™ examination was first administrated in its revised format in June 2011. Revisions were made to all sections of the examination. All revisions were piloted and pre-tested prior to the 2011 administration and stakeholders were informed well in advance of the changes. The purpose of the workshop was to collect recommendations for establishing the BCCE™ examination cut score. The standard setting workshop was held over four days (July 26th – July 29th, 2011) in Athens, Greece, with the participation of judges from Greece and a facilitator from the Hellenic American University.

2. Methodology

This section includes a description of the panel, the materials used, the Benchmarking and Modified Angoff methods, the process implemented during the workshop; and the results from the workshop. The methodology of the workshop is presented in chronological order (i.e. before, during, and after the workshop).

2.1 Selection of Benchmarking and Standard Setting Methods

For the Speaking and Writing sections of the BCCE™ examination, the standard setting method used was based on Chapter 5 of the Manual for Relating Examinations to the Common European Framework of Reference (Council of Europe, 2009). Judges were asked to match each response with a CEFR level.

For the Listening and Grammar, Vocabulary, and Reading (GVR) sections of the BCCE™ examination, a modified Angoff method was used. The Angoff method and its variations have been cited in the literature as the most widely used (Cohen, Kane & Crooks, 1999; Council of Europe, 2009; Kane 1998), and the most thoroughly researched (Brandon, 2004; Cizek & Bunch, 2007; Irwin, Plake, & Impara, 2000) test-centered standard setting methods "designed for use with objective tests and tend[s] to work fairly smoothly with such test[s]" (Kane, 1998: 141). In order to calculate the cut score, three methods are commonly used: (1) the mean (average); (2) the median; and (3) the trimmed mean of the judges' ratings (Zieky, Perie, & Livingston, 2008).

In the modified Angoff method, judges were asked to think of 100 students that barely met the minimum CEFR B1 level criteria (requirements) – that is, 100 students that had just passed the border between CEFR A2 and B1 levels – and to state what proportion of those students would answer each item correctly. In other words, judges entered a probability estimate of how many of those students would get the item correct.

The modified Angoff method was organized in two rounds to allow judges to review their estimates based on *normative information* received after Round 1. Consequently, judges were able to compare their own standards with the rest of the groups' and examination data (Maurer & Alexander, 1992) prior to entering their Round 2 ratings. After Round 2, judges were provided with *impact data* (the percentage of candidates that would pass or fail based on Round 2 recommended cut score) so that they could indicate their confidence in the recommended cut score while completing their evaluation form on the final day.

2.2 Selection of Judges

The recommended number of judges that should be invited to a modified Angoff standard setting workshop is "at least 10 and ideally 15 to 20 judges" (Brandon, 2004: 68). Consequently, the panel of judges consisted of 16 participants (originally 17 judges). The main criteria for judge selection were based on the judges' familiarity with (1) the examinee population; and (2) the intended CEFR test level (Raymond & Reid, 2001). Consequently, all the judges were recruited from Greece and were chosen because the vast majority of the candidates from the June 2011 BCCE™ examination administration were of Greek origin.

The initial invitation sent out to recruit judges specified that the minimum requirements to be considered to serve on the panel were the following:

1. BA holder or its equivalent in English Literature and Linguistics, or a related field
2. A minimum of five years EFL teaching experience
3. A minimum of three years oral examining experience

Judges were asked to fill in a background questionnaire and to attach their CV for consideration. After careful screening of the judges' background questionnaires, a panel of 17 judges meeting the main criteria and the minimum requirements stated above was selected. The judges were sent an e-mail describing the purpose of the standard setting workshop and requesting confirmation of their availability and interest. The judges were also assigned a pre-workshop task of reviewing CEFR scales and online CEFR tasks (see Appendix 1 for e-mail confirmation).

Table 2.1 provides a summary of the judges' background information. One of the judges (J16) had to withdraw during the first day of the workshop for personal reasons and is excluded from the data and the analysis presented in the report.

Table 2.1 Summary of the Judges' Background Information (N = 16)

	Judges' Background Information
1. Gender:	Male **(2);** Female **(14)**
2. Age:	31 – 35 **(3)**; 35+ **(13)**
3. Education:	BA holders **(16);** MA holders **(9);** Other: MA (dissertation pending) **(1)**; RSA holders **(2)**
4. Current Position:	Private English Language Tutors **(6)**; Private School English Language Teachers **(8)**; State School English Language Teachers **(2)**; Administration officers **(2)**; Oral Examiners **(15)**
5. Teaching Experience:	6 – 10 years **(3);** 11 – 15 years **(4);** 16 – 20 years **(4);** 20+ years **(5)**
6. CEFR Levels Taught During Last Academic Year (September 2010 – June 2011) :	Below A1 **(2)**; A1 **(7)**; A2 **(5)**; B1 **(10);** B2 **(9)**; C1 **(12)**; C2 **(8)**
7. CEFR Familiarity:	Not familiar **(1)** A little bit familiar **(1)** Familiar **(10)** Very familiar **(4)**
8. Formal Training with CEFR Descriptors:	None **(7)** Yes **(8)** Other **(1)**
9. Standard Setting Workshop:	Yes **(2)** No **(14)**
10. Oral Examining Experience:	1 examination board only **(3)** At least 2 examination boards **(13)**
11. Written Marking Experience:	None **(7)**; 1 examination board only **(7)**; At least 2 examination boards **(2)**

The panel consisted of teachers who offered private instruction, worked in private language schools, or in state schools. Consequently, the panel represented stakeholders' interest as candidates preparing for the BCCE™ examination will probably have received formal instruction in a private language school, state school or through private instruction. Two of the judges also held administration positions. Only one judge reported no prior familiarity with the CEFR descriptors and one reported little familiarity with the descriptors. Approximately half the panel had received formal training in CEFR descriptors during the Familiarization stage of the BCCE™ examination CEFR linking project.

All judges indicated that they had experience with students across multiple CEFR levels. During the academic year (2010 – 2011), 10 judges taught students at the B1 level, three judges taught students at an adjacent level (A2 or B2), and three judges, who belonged to the original participants in the BCCE™ examination linking project, taught students at more advanced levels (C1 & C2).

All judges were oral examiners for at least one examination board and approximately half the panel worked as raters for at least one examination board. Judges signed a confidentiality agreement on the first day of the workshop (see Appendix 2 for Confidentiality Agreement Form).

2.3 Materials Used for Familiarization and Training Tasks

Prior to the standard setting workshop, material to familiarize judges with the CEFR levels was prepared. CEFR and Dialang scales were separated into "independent meaning units" (Kaftandjieva & Takala, 2002: 107) and judges were asked to sort the descriptors into the six CEFR levels (see Appendix 3 for an example). Despite the fact that the BCCE™ examination aims at a B1 CEFR level, the descriptors used in the familiarization stage did not cover only the B1 level and its two adjacent levels (A2 and B2), but covered all six CEFR levels. As suggested by Kaftandjieva (2010), judges should be provided with all level descriptors, as it is "the judges …who, through their evaluation activity, pose the limitations on the range of levels for the respective competence" (p.46). The descriptors can be viewed as Performance Label Descriptors (PLDs) which judges use "as a critical referent for their judgments" (Cizek, Bunch, & Koons, 2004: 33). Materials used for training were taken from the Council of Europe DVD (2008), the Council of Europe CD (2005), Dialang, and Tanko (2004).

The familiarization tasks were completed on paper or online. Once judges completed their matching activities, they were provided with a key to compare their ratings with the correct CEFR levels. The online tasks were presented as quizzes and judges entered their ratings directly into an online-learning platform (Blackboard Inc., 2010). By using this medium, judges received immediate feedback on their ratings and the facilitator could monitor in real time who had finished each activity and how many of the statements were correctly ranked (see Appendix 3 for examples of familiarization paper and online tasks, judge feedback, and facilitator's view of results). At the end of each task, a discussion took place so that a shared understanding was reached on the correct matching of each descriptor.

Table 2.2 illustrates the materials used during the Benchmarking and Standard Setting for the familiarization and training stages. A total of 241 descriptors and 44 items were used to familiarize and train the judges with the CEFR levels.

***Table 2.2** Familiarization & Training Materials*

Section	No. of Descriptors	Familiarization Stage	No. of Items	Training Stage
Overview	9	Table A1: Salient Characteristics: Interaction & Production (CEFR Section 3.6, simplified)		
Speaking	51	Overall Speaking Table C1: Global Oral Assessment Scale Table C2: Oral Assessment Criteria Grid	14	(Council of Europe, DVD, 2008)
Writing	53	Overall Writing Interaction Correspondence Overall Written Production Reports and Essays Table C4 – Written Assessment Criteria Grid	9	Tanko (2004)
Grammar	32	DIALANG Grammar Scales	3	BCCE™ examination previously administered grammar items
Vocabulary	28	DIALANG Vocabulary Scales	-	-
Reading	32	Overall Reading Comprehension Reading Correspondence Reading for Orientation Reading for Information and Argument	12	DIALANG (Council of Europe, CD, 2005)
Listening	36	Overall Listening Comprehension Understanding Interaction Between Native Speakers Listening as a Member of a Live Audience Listening to Announcements and Instructions Listening to Audio Media and Recordings	6	DIALANG (Council of Europe, CD, 2005)
Total	241		44	

The standard setting workshop began with a presentation on the BCCE™ examination CEFR linking project and an orientation on the purpose of the standard setting workshop (see Appendix 4 for the workshop agenda).

2.4 Judgment Procedure for Standardization Training and Benchmarking: Speaking Section

The procedure for Standardization Training and Benchmarking of the Speaking section was based on Chapter 5 of the Manual (Council of Europe, 2009) and consisted of the following steps: (1) Familiarization; (2) Working with Standardized Samples; and (3) Benchmarking Local Samples.

Step 1: Familiarization

This step entailed working with the CEFR descriptors and matching the descriptors with their corresponding CEFR levels. A total of 51 CEFR descriptors were used during this step. The descriptors were taken from the following CEFR scales: (1) Overall Speaking; (2) Table C1: Global Oral Assessment Scale; (3) Table C2: Oral Assessment Criteria Grid. Judges were asked to rank order the descriptors from A1 to C2. At the end of each familiarization task, judges were given feedback on their answers and any misplaced descriptors were discussed so that all judges came to an agreement on the correct placing of each descriptor. Hard copies of all the scales used during this step were provided to judges to use during the next steps of the judgment process.

Step 2: Working with Standardized Speaking Samples

This step was divided in three phases (illustration, controlled practice, and freer practice) as suggested in the CEFR Manual (Council of Europe, 2009). A total of seven pairs of students were evaluated during this step. Judges were asked to match each performance with a CEFR level (A1, A2, A2+ B1, B1+ B2, B2+, C1, C2) (see Appendix 5 for an example of the rating form). A total of 14 students (six students during illustration, four students during controlled practice, and four students during freer practice) were selected for standardization. Since most judges worked as oral examiners for more than one examination board, the facilitator deemed further training with the CEFR descriptors and benchmarked samples necessary in case panelists were previously "influenced by local institutional standards intended to be at CEFR levels and criterion descriptors for them or locally produced variants of CEFR descriptors" (Council of Europe, 2009: 17).

After each phase, judges discussed their individual ratings in groups of three or four and then a panel discussion on each sample took place in which the facilitator provided the correct CEFR level and its rationale according to the CEFR documentation on each sample. Individual judgments were collected for processing after each phase.

Step 3: Benchmarking Local Speaking Samples

This step was also divided into three phases and each phase consisted of two rounds. Three audio samples were played during each phase and judges were asked to listen to each audio recording and to enter a score from 1 to 4 corresponding to the following criteria: 1= A2; 2 = A2+; 3 = B1; and 4= B1+ which were recoded for analysis as: 1 = Below minimum B1 level; 2 = At minimum B1 level; and 3 = Above minimum B1 level. The recoding took place as some judges expressed their unease with working with the CEFR plus (+) levels in phase one and avoided using them.

After Round 1, judges were asked to discuss their ratings in pairs and then a panel discussion followed in which judges provided a rationale for their rankings. At the end of Round 1, judges were given feedback on their ratings. Judges were then provided with an opportunity to change their rankings in Round 2.

The same procedure was followed for the other two phases of the Benchmarking session. On the last day of the workshop, judges were provided with their recommended Speaking section cut score so that they could comment on it while completing the end of workshop evaluation form.

2.5 Judgment Procedure for Standardization Training and Benchmarking: Writing Section

The procedure for Standardization Training and Benchmarking of the Writing section was also based on Chapter 5 of the Manual (Council of Europe, 2009) and consisted of the following steps: (1) Familiarization; (2) Working with Standardized Samples; and (3) Benchmarking Local Samples.

Step 1: Familiarization

This step entailed working with the CEFR descriptors and matching the descriptors with their corresponding CEFR levels. A total of 53 descriptors were selected from the following CEFR scales: (1) Overall Writing Interaction Correspondence; (2) Overall Written Production; (3) Reports and Essays; and (4) Table C4 – Written Assessment Criteria Grid. At the end of each familiarization task, judges received feedback on their answers and any misplaced descriptors were discussed so that all judges came to an agreement on the correct placing of each descriptor. Hard copies of all the scales used during this step were provided to judges to refer to during the next steps of the judgment process.

Step 2: Working with Standardized Writing Samples

Originally, six samples were used from the Council of Europe CD (2005); however, the samples came from the Cambridge ESOL suite and nearly all judges chose the correct CEFR level based on their experience with the examination suite. Following Papageorgiou (2010a), written samples for training purposes were taken from Tanko (2004). A total of nine samples (three tasks) were used for the training step. Task 1 (four samples) was a transactional letter; Task 2 (two samples) was an informal letter; and Task 3 (three samples) was an article. Judges were asked to match each sample with one of the following CEFR levels: A2, B1, or B2. It should be noted that there were no correct answers provided to judges concerning the training items. Tanko (2004) claims that the samples had been benchmarked and are between a B1 level and a B2 level inclusive. Nonetheless, the samples allowed judges to discuss their ratings and to reevaluate their interpretation of writing proficiency at CEFR levels A2, B1, and B2.

Step 3: Benchmarking Local Writing Samples

This step was divided in three phases (illustration, controlled practice, and freer practice) as suggested in the Manual (Council of Europe, 2009). For Task 1 (E-mail 1), phase 1 consisted of five scripts, phase 2 consisted of 10 scripts, and phase 3 consisted of nine scripts. The three phases were repeated for Task 2 (E-mail 2): phase 1 consisted of five scripts, phase 2 consisted of 10 scripts, and phase 3 consisted of six scripts. A total of 45 (24 for Task 1 and 21 for Task 2) local writing scripts were used.

Judges were asked to indicate at which CEFR level (A2, B1, or B2) they considered each writing sample to be. It should be noted that judges were not asked to use the BCCE™ Writing Scoring Rubrics, as an additional number of hours would have been required for training judges to use the rubrics.

Each phase consisted of two rounds. Judges entered their Round 1 ratings in their writing script booklet and then discussed their ratings in groups of three and four. A panel discussion followed after which judges were given the opportunity to review their initial judgments of each script. Judges entered their Round 2 ratings for each script on a separate sheet (see Appendix 6 for the Benchmarking Writing section rating form).

The following day, judges were provided with their recommended Writing section cut score so that they could comment on it when completing their end of workshop evaluation form.

2.6 Judgment Procedure for Standard Setting: GVR Section

The procedure for the standard setting of the GVR sections consisted of the following steps: (1) familiarization; (2) training; (3) GVR section timed; (4) standard setting method training; (5) Round 1; and (6) Round 2.

Step 1: Familiarization

For the GVR section, this step entailed working with CEFR and Dialang Descriptors and matching the descriptors with their corresponding CEFR levels. A total of 92 descriptors were selected from the following CEFR and Dialang scales: (1) Overall Reading Comprehension; (2) Reading Correspondence; (3) Reading for Orientation: (4) Reading for Information and Argument; (5) Dialang Grammar; and (6) Dialang Vocabulary. At the end of each familiarization task, judges were given feedback on their answers and any misplaced descriptors were discussed so that all judges came to an agreement on the correct placing of each descriptor. Hard copies of all the scales used during this step were provided to judges to refer to during the next steps of the judgment process.

Step 2: Training
For the GVR section, this step entailed working with 12 benchmarked CEFR reading texts and items from the Council of Europe CD (2005). Judges were asked what the minimum CEFR level required of a candidate to be able to answer the item correctly was.

At the end of this step, judges received feedback on their answers and a discussion took place on the misplaced items.

Step 3: GVR section timed
Judges were asked to take the GVR section test under timed conditions and were then provided with a key to compare their answers.

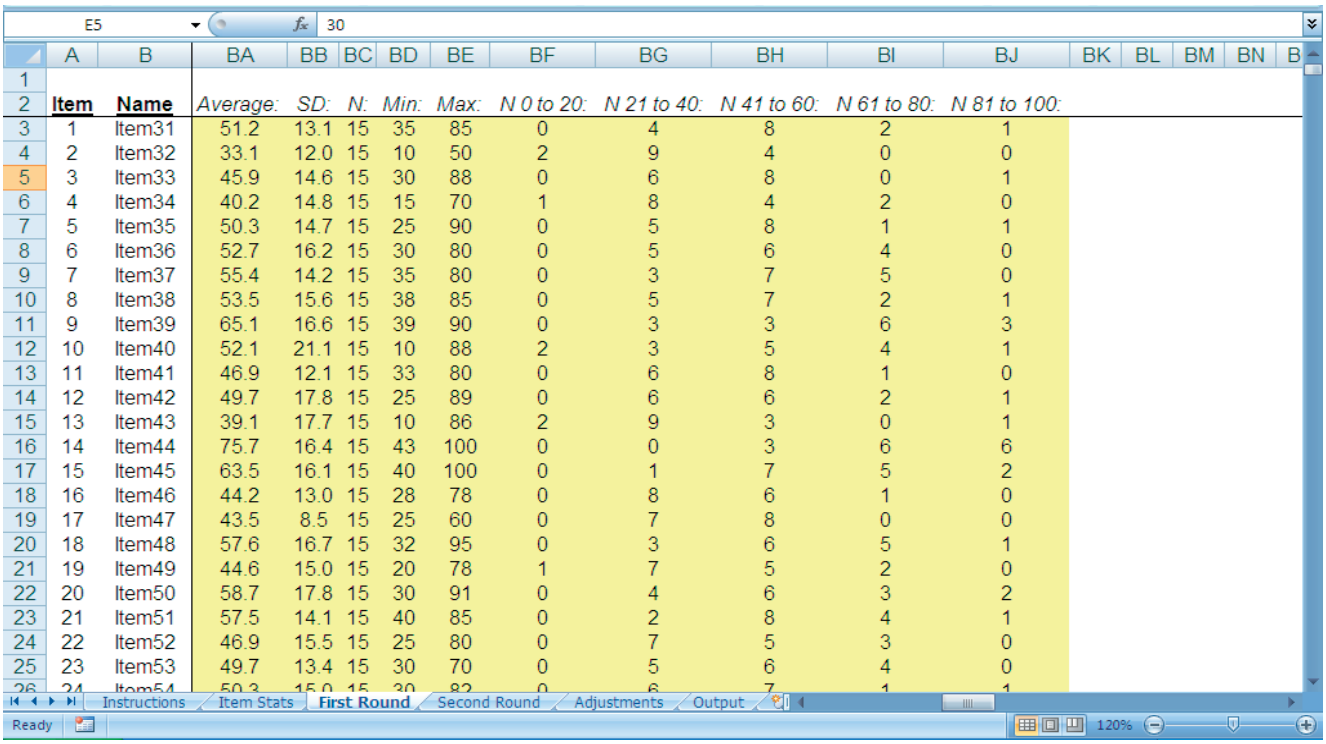
Step 4: Standard setting method training
The Modified Angoff method was explained to the judges. They were instructed to think of 100 students who barely met the minimum B1 level criteria (requirements) and to record what proportion of those students would answer each item correctly. In other words, judges were asked to record how many of those 100 students would get the item correct.

Prior to the judges making their Round 1 ratings for the GVR section, they were provided with an opportunity to become familiar with the rating process by providing judgments on three grammar items from a previous administration. No vocabulary items were used for training as the revised BCCE™ vocabulary part of the GVR section is different from the vocabulary part of previous administrations. Each item was presented one at a time and after judges completed their performance estimates for an item, they were asked to share their estimates with the rest of the panel and to provide reasons for their estimates. A discussion revolved around the judges' reasoning. At the end of this step, judges were asked whether they were ready to begin with their Round 1 ratings.

Step 5: Round 1
During Round 1, judges were instructed to work independently and enter their Round 1 ratings into a Microsoft Excel® worksheet. At the end of Round 1, all of the data were entered into the Angoff Analysis Tool (AAT) - a specifically designed Microsoft Excel® workbook for Angoff ratings (Assessment Systems Corporation, 2009) - and projected to the judges. The judges were presented with *normative information* illustrating their ratings on each item and their individual cut score for Round 1. To facilitate Round 1 discussion, descriptive statistics for each item were presented.

Figure 2 is an example of the item statistics shown to judges during GVR Round 1 discussion. The first column is the item number, the second column shows the item ID, the third column illustrates the average rating of the item, the fourth column shows the standard deviation of the judges' ratings, the fifth columns shows the number of judges (note: J13 was absent). The sixth and seventh columns show the minimum and maximum rating for the item respectively. The next five columns display how many judges entered a rating within one of the five bands: (1) 0% - 20%; (2) 21% - 40%; (3) 41% - 60%, (4) 61% - 80%; and (5) 81% - 100%. When an item spread across at least three different bands, judges from each band were asked to share their rationale for their ratings.

***Figure 2** GVR Round 1 Discussion: Descriptive Statistics*



Item	Name	Average:	SD:	N:	Min:	Max:	N 0 to 20:	N 21 to 40:	N 41 to 60:	N 61 to 80:	N 81 to 100:
1	Item31	51.2	13.1	15	35	85	0	4	8	2	1
2	Item32	33.1	12.0	15	10	50	2	9	4	0	0
3	Item33	45.9	14.6	15	30	88	0	6	8	0	1
4	Item34	40.2	14.8	15	15	70	1	8	4	2	0
5	Item35	50.3	14.7	15	25	90	0	5	8	1	1
6	Item36	52.7	16.2	15	30	80	0	5	6	4	0
7	Item37	55.4	14.2	15	35	80	0	3	7	5	0
8	Item38	53.5	15.6	15	38	85	0	5	7	2	1
9	Item39	65.1	16.6	15	39	90	0	3	3	6	3
10	Item40	52.1	21.1	15	10	88	2	3	5	4	1
11	Item41	46.9	12.1	15	33	80	0	6	8	1	0
12	Item42	49.7	17.8	15	25	89	0	6	6	2	1
13	Item43	39.1	17.7	15	10	86	2	9	3	0	1
14	Item44	75.7	16.4	15	43	100	0	0	3	6	6
15	Item45	63.5	16.1	15	40	100	0	1	7	5	2
16	Item46	44.2	13.0	15	28	78	0	8	6	1	0
17	Item47	43.5	8.5	15	25	60	0	7	8	0	0
18	Item48	57.6	16.7	15	32	95	0	3	6	5	1
19	Item49	44.6	15.0	15	20	78	1	7	5	2	0
20	Item50	58.7	17.8	15	30	91	0	4	6	3	2
21	Item51	57.5	14.1	15	40	85	0	2	8	4	1
22	Item52	46.9	15.5	15	25	80	0	7	5	3	0
23	Item53	49.7	13.4	15	30	70	0	5	6	4	0

For example, item 1 (item ID = 31) had an average rating of 51.2 and a standard deviation of 13.1. The minimum rating was 35 while the maximum rating was 85. Four judges entered a rating between 21% and 40%, eight judges entered a rating between 41% and 60%, two judges entered a rating between 61% and 80% and one judge entered a rating between 81% and 100%. Item 1 spread across four different bands and judges from each band were asked to share with the rest of the panel their rationale for their rating.

At the end of the discussion, judges were shown their inter-rater reliability and their Round 1 recommended cut score. Following Round 1 discussion, judges were given *reality information* in the form of p-values and point-biserial correlations (i.e. discrimination indices) for each item. A discussion took place on how to interpret p-values and point-biserial correlations. Judges were reminded that the p-values and point-biserial correlations that they were provided with were based on the entire testing population and that they were entering ratings for the minimum competent candidates. Judges were also provided with an extract from Iteman 4.0 user's manual (Thompson & Guyer, 2010) explaining how to interpret the *reality information* (see Appendix 7 for extract).

Step 6: Round 2

Judges were then advised to consider all information presented, to reevaluate their first estimates, and to repeat the Round 1 process. It was emphasized that they were not obliged to change any estimates should they wish not to. Another Microsoft Excel® worksheet was distributed in which the judges entered their second ratings. The worksheet also contained the p-values and the point-biserial values of each item.

At the end of Round 2, judges were shown *normative information*, *impact* or *consequential data* (pass rate based on Round 2 recommended cut score), the groups' inter-rater reliability (the degree to which their ratings agreed), and Round 2 recommended cut score.

2.7 Judgment Procedure for Standard Setting: Listening Section

The procedure for the standard setting of the Listening section was similar to that of the GVR section and consisted of the following steps: (1) familiarization; (2) training; (3) listening section timed; (4) Round 1; and (5) Round 2. As judges had received training in the standard setting method when the standard setting for the GVR section took place, no additional training took place.

Step 1: Familiarization

This step entailed working with 36 descriptors taken from the following CEFR scales: (1) Overall Listening Comprehension; (2) Understanding Interaction Between Native Speakers; (3) Listening as a Member of a Live Audience; (4) Listening to Announcements and Instructions; and (4) Listening to Audio Media and Recordings. Hard copies of all the scales used during this step were provided to judges to refer to during the next steps of the judgment process.

Step 2: Training

This step entailed working with six benchmarked CEFR texts and items taken from the Council of Europe CD (2005). Judges were asked what the minimum CEFR level required of a candidate to be able to answer the item correctly was.

At the end of this step, judges received feedback on their answers and a discussion took place on the misplaced items.

Step 3: Listening section timed

Judges were asked to take the Listening section under timed conditions and were then provided with a key to compare their answers.

Step 4: Round 1

Judges were asked to think of 100 students that barely met the minimum B1 level criteria (requirements) and to record what proportion of those students would answer each item correctly. In other words, judges were asked to record how many of those 100 students would get the item correct.

The Listening section was played again and judges were instructed to work independently and enter their Round 1 estimates into a Microsoft Excel® worksheet (see Appendix 8 for Round 1 worksheet). At the end of Round 1, all of the judges' data were entered into the Angoff Analysis Tool (AAT) Excel® workbook (Assessment Systems Corporation, 2009) and projected to the judges. The judges were presented with *normative information* illustrating their ratings on each item and their individual cut score for Round 1. Round 1 discussion was conducted in the same way as it was done during the Round 1 discussion of the GVR section. Judges were presented with their responses in five bands. When an item spread across at least three different bands, judges from each band were asked to share their rationale for their ratings. At the end of the discussion, judges were shown their inter-rater reliability and their Round 1 recommended cut score. Following Round 1 discussion, judges were provided with *reality information* in the form of p-values and point-biserial correlations for each item.

Step 6: Round 2

Judges were then advised to consider all information presented and to reevaluate their first estimates. Judges were reminded that they were not obliged to change any of their original estimates should they wish not to. Another Microsoft Excel® worksheet (see Appendix 8 for Round 2 worksheet) was distributed in which the

judges entered their second ratings. The worksheet also contained the p-values and the point-biserial values for each item. The Listening section was not played a third time. At the end of Round 2, judges were shown *normative information*, *impact data*, the groups' inter-rater reliability, and their Round 2 recommended cut score.

3. Results of CEFR Familiarization and Training Activities

Prior to analyzing the cut scores, the CEFR familiarization and training activities will be discussed so that the judges' training with the CEFR levels can be established. Before recommending a cut score, judges should display a good understanding of the CEFR levels and should be able to rank order the CEFR descriptors with the correct level (Papageorgiou, 2010b). Otherwise, if judges cannot display a reasonable understanding of the CEFR levels and descriptors, the validity of their recommended cut scores may be questioned.

Following the procedure adopted by Papageorgiou (2010b), each judge's familiarization and training tasks were analyzed in terms of: (1) the number of correctly placed items; (2) the correlations between item placement and correct CEFR level placement; and (3) a comparison between the judge's task mean and its CEFR task mean. For data analysis, the following codes were used: A1 = 1; A2 = 2; B1 = 3; B2 = 4; C1 = 5; and C2 = 6.

Table 3.1 to Table 3.9 present the judges' rankings during the familiarization and training tasks. The first row displays the number of descriptors correctly placed. The second row presents the spearman correlation between the judges' rankings and the correct CEFR levels, while the last row illustrates the judge's task mean. The judge's task mean can be compared with the CEFR task mean (indicated in the parenthesis of each table's caption). A comparison of a judge's task mean with the CEFR task mean may illustrate a judge's tendency of assigning descriptors or items to a higher or lower level than the correct CEFR level.

For example, in Table 3.1, J1's task mean was 3.33 while the CEFR task mean was 3.49. J1 placed the 51 descriptors as follows: 11 at A1, nine at A2, seven at B1, eight at B2, eight at C1, and eight at C2. J1's task mean was 3.33 as illustrated below:

$$\frac{[(11 \times 1) + (9 \times 2) + (7 \times 3) + (8 \times 4) + (8 \times 5) + (8 \times 6)]}{51} = 3.33$$

The correct CEFR placement of the same descriptors is as follows: eight at A1, nine at A2, nine at B1, eight at B2, nine at C1, and eight at C2. The CEFR task mean was 3.49 as illustrated below:

$$\frac{[(8 \times 1) + (9 \times 2) + (9 \times 3) + (8 \times 4) + (9 \times 5) + (8 \times 6)]}{51} = 3.49$$

The difference between J1's task mean (3.33) and the CEFR task mean (3.49) indicates that J1 assigned some descriptors at a lower level than their actual CEFR level. When a judge's task mean is less than the CEFR task mean, then that judge expresses some tendency of severity. Conversely, when a judge's task mean is greater than the CEFR task mean, then that judge expresses some tendency of leniency. For example, J15's task mean was 3.57 while the CEFR task mean was 3.49. This indicates that J15 may have expressed some tendency of leniency by assigning some descriptors at a higher level than their actual CEFR level.

As the writing training tasks used were taken from Tanko (2004) and do not have a key, no Writing training table could be created.

***Table 3.1** Speaking Familiarization Task Results (51 descriptors, CEFR task mean 3.49)*

	J1	J2	J3	J4	J5	J6	J7	J8	J9	J10	J11	J12	J13	J14	J15	J17
Correct	35	40	42	39	42	28	31	47	39	44	21	35	34	39	43	40
Spearman	0.95	0.95	0.97	0.95	0.97	0.85	0.84	0.98	0.95	0.98	0.75	0.92	0.94	0.91	0.95	0.93
Task Mean	3.33	3.49	3.43	3.51	3.43	3.35	3.43	3.49	3.47	3.43	3.37	3.49	3.43	3.57	3.57	3.41

During the Speaking Familiarization tasks, 51 descriptors were used and the judges' correct placement of descriptors ranged from 21 to 47. The correlation between the correct placement and the correct levels ranged from .75 to .98 while the judges' task mean ranged from 3.33 to 3.57. It seemed that the judges expressed some tendency of severity as indicated by the fact that 10 judges had a mean less than the CEFR task mean (3.49).

***Table 3.2** Speaking Training Task Results (14 items, CEFR task mean 3.14)*

	J1	J2	J3	J4	J5	J6	J7	J8	J9	J10	J11	J12	J13	J14	J15	J17
Correct	6	10	12	7	12	6	7	11	7	9	10	9	10	9	10	10
Spearman	0.75	0.85	0.95	0.84	0.95	0.86	0.69	0.95	0.69	0.82	0.90	0.86	0.91	0.84	0.91	0.91
Task Mean	3.36	3.29	3.29	3.64	3.29	2.93	3.36	3.43	3.36	3.21	3.14	3.21	3.29	3.21	3.29	3.29

During the Speaking Training tasks, 14 items were used and the judges' correct placement of items ranged from 6 to 12. The correlation between the correct placement and the correct levels ranged from .69 to .95 while the judges' task mean ranged from 2.93 to 3.64. It seemed that the judges expressed some tendency of leniency as indicated by the fact that 14 judges had a mean greater than the CEFR task mean (3.14).

***Table 3.3** Writing Familiarization Task Results (53 descriptors, CEFR task mean 3.55)*

	J1	J2	J3	J4	J5	J6	J7	J8	J9	J10	J11	J12	J13	J14	J15	J17
Correct	33	40	52	29	52	29	35	53	43	36	28	33	36	42	41	37
Spearman	0.86	0.94	0.98	0.75	1.00	0.86	0.85	1.00	0.95	0.90	0.77	0.89	0.93	0.95	0.95	0.88
Task Mean	3.42	3.40	3.51	3.36	3.53	3.60	3.74	3.55	3.60	3.66	3.40	3.64	3.70	3.43	3.66	3.43

During the Writing Familiarization tasks, 53 descriptors were used and the judges' correct placement of descriptors ranged from 28 to 53. The correlation between the correct placement and the correct levels ranged from .75 to 1.0 while the judges' task mean ranged from 3.40 to 3.74. It seemed that half the judges expressed some tendency of severity as indicated by the fact that eight judges had a mean less than the CEFR task mean (3.55).

***Table 3.4** Grammar Familiarization Task Results (32 descriptors, CEFR task mean 3.13)*

	J1	J2	J3	J4	J5	J6	J7	J8	J9	J10	J11	J12	J13	J14	J15	J17
Correct	22	16	17	10	18	17	10	17	20	11	16	17	18	14	18	17
Spearman	0.90	0.93	0.83	0.65	0.71	0.66	0.63	0.86	0.86	0.73	0.72	0.69	0.87	0.76	0.87	0.83
Task Mean	3.16	2.69	2.81	3.56	2.75	2.28	2.38	2.78	2.94	2.34	2.56	2.88	2.97	2.88	2.78	2.69

During the Grammar Familiarization tasks, 32 descriptors were used and the judges' correct placement of descriptors ranged from 10 to 22. The correlation between the correct placement and the correct levels ranged from .63 to .93 while the judges' task mean ranged from 2.28 to 3.56. It seemed that the judges expressed some tendency of severity as indicated by the fact that 14 judges had a mean less than the CEFR task mean (3.13).

***Table 3.5** Vocabulary Familiarization Task Results (28 descriptors, CEFR task mean 3.43)*

	J1	J2	J3	J4	J5	J6	J7	J8	J9	J10	J11	J12	J13	J14	J15	J17
Correct	16	17	17	8	19	15	18	13	17	12	11	12	12	15	14	14
Spearman	0.89	0.93	0.87	0.79	0.95	0.88	0.91	0.93	0.92	0.87	0.69	0.94	0.85	0.94	0.94	0.84
Task Mean	3.43	3.64	3.71	4.07	3.61	3.61	3.32	3.68	3.71	3.04	3.39	4.18	3.71	3.93	3.96	3.07

During the Vocabulary Familiarization tasks, 28 descriptors were used and the judges' correct placement of descriptors ranged from 8 to 19. The correlation between the correct placement and the correct levels ranged from .69 to .95 while the judges' mean level placement ranged from 3.04 to 4.18. It seemed that half the judges expressed some tendency of leniency as indicated by the fact that 11 judges had a mean greater than the CEFR task mean (3.43).

***Table 3.6** Reading Familiarization Task Results (32 descriptors, CEFR task mean 3.09)*

	J1	J2	J3	J4	J5	J6	J7	J8	J9	J10	J11	J12	J13	J14	J15	J17
Correct	18	20	26	13	30	10	20	23	16	19	17	25	21	20	20	21
Spearman	0.93	0.92	0.97	0.60	0.99	0.77	0.92	0.93	0.90	0.88	0.77	0.94	0.92	0.86	0.92	0.86
Task Mean	3.22	3.09	3.16	3.25	3.09	3.56	3.16	3.00	3.59	3.16	3.22	3.19	3.06	3.28	3.41	3.22

During the Reading Familiarization tasks, 32 descriptors were used and the judges' correct placement of descriptors ranged from 10 to 30. The correlation between the correct placement and the correct levels ranged from .60 to .99 while the judges' mean level placement ranged from 3.00 to 3.59. It seemed that the judges expressed some tendency of leniency as indicated by the fact that 12 judges had a mean greater than CEFR task mean (3.09).

***Table 3.7** Reading Training Task Results (12 items, CEFR task mean 3.17)*

	J1	J2	J3	J4	J5	J6	J7	J8	J9	J10	J11	J12	J13	J14	J15	J17
Correct	10	11	10	10	11	8	10	12	11	11	10	10	11	11	10	11
Spearman	0.95	0.99	0.96	0.96	0.99	0.90	0.90	1.00	0.99	0.97	0.92	0.91	0.99	0.99	0.96	0.99
Task Mean	3.33	3.25	3.33	3.33	3.25	3.50	3.33	3.17	3.25	3.25	3.33	3.42	3.25	3.25	3.33	3.25

During the Reading Training tasks, 12 items were used and the judges' correct placement of items ranged from 8 to 12. The correlation between the correct placement and the correct levels ranged from .90 to 1.00 while the judges' mean level placement ranged from 3.17 to 3.50. It seemed that the judges expressed some tendency of leniency as indicated by the fact that nearly all of the judges (15) judges had a mean greater than the CEFR task (3.17).

***Table 3.8** Listening Familiarization Task Results (36 descriptors, CEFR task mean 3.56)*

	J1	J2	J3	J4	J5	J6	J7	J8	J9	J10	J11	J12	J14	J15	J17
Correct	19	29	35	31	35	20	23	35	23	21	18	17	19	35	26
Spearman	0.91	0.94	0.99	0.95	0.99	0.90	0.88	0.99	0.91	0.87	0.78	0.84	0.85	0.99	0.93
Task Mean	3.53	3.53	3.58	3.56	3.58	3.78	3.44	3.58	3.81	3.36	3.61	3.69	3.61	3.58	3.56

During the Listening Familiarization tasks, 36 descriptors were used and the judges' correct placement of descriptors ranged from 17 to 35. The correlation between the correct placement and the correct levels ranged from .78 to .99 while the judges' mean level placement ranged from 3.36 to 3.78. It seemed that about half the judges expressed some tendency of leniency as indicated by the fact that nine judges had a mean greater than the CEFR task mean (3.56).

***Table 3.9** Listening Training Task Results (6 items, CEFR task mean 2.83)*

	J1	J2	J3	J4	J5	J6	J7	J8	J9	J10	J11	J12	J14	J15	J17
Correct	4	5	3	4	3	3	4	3	4	3	3	4	5	3	4
Spearman	0.81	1.00	0.95	0.90	0.95	0.86	0.95	0.95	0.77	0.42	0.95	0.94	0.89	0.95	1.00
Task Mean	2.67	2.83	3.17	3.00	3.17	2.83	2.33	3.17	3.00	2.50	3.17	3.17	2.83	3.17	3.00

During the Listening Training tasks, six items were used and the judges' correct placement of items ranged from 3 to 5. The correlation between the correct placement and the correct levels ranged from .42 to 1.00 while the judges' mean level placement ranged from 2.50 to 3.17. It seemed that half the judges expressed some tendency of leniency as indicated by the fact that eight judges had a mean greater than the CEFR task mean (2.83).

The fact that the descriptors in the familiarization tasks were separated into independent units made the task of correct placement more challenging. Nonetheless, the high correlations in the familiarization and the training tasks suggest that the judges had a good understanding of the CEFR levels and "how the descriptors progress from lower to higher levels"(Papageorgiou, 2010b: 4). Overall, the judges seemed to exhibit some tendency of leniency during most familiarization and training tasks. This issue was addressed after each task and descriptors and items were discussed until an understanding of their correct placement was achieved.

As cut scores are calculated on group estimates, inter-rater reliability estimates were calculated on the judges' familiarization and training tasks. Tables 3.10 and 3.11 display the inter-rater reliability estimates for the familiarization and training tasks. As "there is no single perfect statistical index for the estimation of inter-rater reliability" (Kaftandjieva & Takala, 2002: 111), both consistency and consensus inter-rater reliability indices are reported.

The first row displays the rater agreement index (Burry-Stock et al., 1996), which shows consistency agreement amongst raters. The rater agreement index (RAI) is "[the] average rater agreement over all descriptor units" (Kaftandjieva & Takala, 2002: 112). A RAI of 1 implies perfect agreement amongst raters while a RAI of 0 implies no agreement amongst raters (Burry-Stock et al., 1996). The second row shows the groups' Cronbach's alpha, which is an internal consistency estimate. A high alpha estimate implies that judge ratings measure a common dimension. The third row shows intraclass correlation, ICC, (McGraw & Wong, 1996: Shrout & Fleiss, 1979). An extremely high intraclass correlation (close to 1) suggests that raters have achieved an excellent inter-rater reliability (Stemler & Tsai, 2008). The model used to calculate the ICC was the two-way mixed model, average measures for exact agreement. Cronbach's alpha and ICC are consensus inter-rater agreement indices. ICC and Cronbach's alpha are two "of the most frequently used indicators of internal consistency" (Kaftandjieva, 2010: 96).

***Table 3.10** Inter-rater Reliability Indices for Familiarization Tasks*

	Speaking	Writing	Grammar	Vocabulary	Reading	Listening
RAI	.93	.92	.89	.88	.91	.92
Alpha	.99	.99	.98	.98	.98	.99
ICC	.99	.99	.97	.98	.98	.99

***Table 3.11** Inter-rater Reliability Indices for Training Tasks*

	Speaking	Writing	Reading	Listening
RAI	.92	.95	.97	.92
Alpha	.98	.97	.99	.99
ICC	.98	.98	.99	.99

For the familiarization tasks, the RAI for each section ranged from .88 for the Vocabulary task to .93 for the Speaking task. The Cronbach's alpha estimate ranged from .98 for the Grammar, Vocabulary, and Reading tasks to .99 for the Speaking, Writing, and Listening tasks. The ICC estimate ranged from .97 for the Grammar task to .99 for the Speaking, Writing, and Listening tasks. For the training tasks, the RAI for each section ranged from .92 for the Speaking and Listening tasks to .97 for the Reading task. The Cronbach's alpha estimate ranged from .97 for the Writing task to .99 for the Reading and Listening tasks. The ICC estimate ranged from .98 for the Speaking and Writing tasks to .99 for the Reading and Listening tasks.

Reliability estimates should be at least .80 to "reflect good dependability of scores" (Hyot, 2010: 152). Examining tables 3.10 & 3.11, the high inter-rater reliability coefficients suggest that the familiarization and training tasks were successful in training judges to differentiate between the different CEFR levels and that the judges "were applying similar constructs when they sorted the descriptor units into six ordered piles" (Kaftandjieva & Takala, 2002: 112) during the familiarization and training tasks.

4. Cut Score Results and Validity Evidence

4.1 Cut Score Internal Validation

Standard setting workshops are evaluated in terms of three elements as illustrated in Table 4.1. The first three elements of Procedural validity were presented in the methodology section of this report and the fourth element (feedback) is addressed towards the end of this section. This section will be concerned with presenting evidence of internal validation and a discussion of future external validity evidence.

***Table 4.1** Standard Setting Evaluation Elements*

Procedural	Internal	External
Explicitness	Consistency within Method	Comparisons to other Standard Setting Methods
Practicability	Intraparticipant Consistency	Comparisons to other Sources of Information
Implementation of procedures	Interparticipant Consistency	Reasonableness of performance levels
Panelist feedback	Decision Consistency	
Documentation	Other Measures	

Source: (Hambleton & Pitoniak, 2006)

Sections 4.2 through 4.6 present the cut score judgments for each section. Both Round 1 and Round 2 estimates and their standard deviations are reported to show the degree to which ratings change across rounds (interparticipant consistency).

4.2 Judgments of Speaking Section

Table 4.2 displays both Round 1 and Round 2 judges' mean scores for each speaking item. The first column displays the item ID. The second column displays the judges' Round 1 mean score for each item and the third column displays the judges' Round 2 mean score for each item. The original ratings of 1= A2; 2 = A2+; 3 = B1; and 4= B1+ were recoded for analysis as: 1 = Below minimum B1 level; 2 = At minimum B1 level; and 3 = Above minimum B1 level. The recoding took place as some judges expressed some unease with using CEFR plus levels and avoided using them. The standard deviation of the judges' scoring for each item and round is reported in parenthesis.

***Table 4.2** Item Judgments for the Speaking Section (N=9)*

	Round 1 mean	(SD)	Round 2 mean	(SD)
SL-1	1.00	(.00)	1.00	(.00)
SL-2	1.06	(.25)	1.07	(.25)
SL-3	2.00	(.52)	2.00	(.44)
SL-4	2.63	(.50)	2.66	(.50)
SL-5	1.88	(.62)	1.93	(.57)
SL-6	1.25	(.45)	1.27	(.25)
SL-7	1.00	(.00)	1.00	(.00)
SL-8	2.94	(.25)	2.93	(.25)
SL-9	1.81	(.40)	1.86	(.40)

An examination of the data reveals that the standard deviation of each item either decreased or remained the same between the two rounds. While a decrease in the standard deviation of each item between rounds is attributed to an increase in consensus amongst the judges since "the lower the standard deviation, the more agreement there is among panelists" (Hambleton & Pitoniak, 2006: 460), the overall rounded mean of each item remains the same (Hambleton, Pitoniak & Copella , 2012).

The judges' CEFR mean scores for each item were rounded to the nearest integer and then cross tabulated with their respective benchmarked scores for those items. Tables 4.3 to 4.4 show the cross tabulation of the judges' mean CEFR scores with the benchmarked scores for Rounds 1 and 2. The first row shows the official benchmarked scores and the first column shows the CEFR level the judges matched the item to. A score of 1 indicates that the item is below a minimum B1 level, a score of 2 indicates that an item is at a minimum B1 level, and a score of 3 indicates that an item is above minimum B1 level.

***Table 4.3** Round 1 Cut Score Judgment for the Speaking Section*

	Benchmarked Speaking Scores							
	9	10	11	12	14	17	20	Total
1 (Below Minimum B1)	1	1	2	0	0	0	0	4
2 (At Minimum B1)	0	0	0	2	1	0	0	3
3 (Above Minimum B1)	0	0	0	0	0	1	1	2
Total	1	1	2	2	1	1	1	9

***Table 4.4** Round 2 Cut Score Judgment for the Speaking Section*

	Benchmarked Speaking Scores							
	9	10	11	12	14	17	20	Total
1 (Below Minimum B1)	1	1	2	0	0	0	0	4
2 (At Minimum B1)	0	0	0	2	1	0	0	3
3 (Above Minimum B1)	0	0	0	0	0	1	1	2
Total	1	1	2	2	1	1	1	9

Tables 4.3 to 4.4 illustrate that the minimum score (shaded cells) for an item to be considered at a minimum B1 level was 12 for both Round 1 and Round 2.

4.3 Judgments of Writing Section

Table 4.5 displays the judges' mean scores for each writing script. Columns 1 and 4 display the items ID code. Columns 2 and 5 display the judges' mean score for each item for Round 1, while Columns 3 and 6 display the judges' mean scores for Round 2. A score of 1 indicates that a script is at an A2 level, a score of 2 indicates that a script is at a B1 level, and a score of 3 indicates that a script is at a B2 level.

***Table 4.5** Item Judgment for the Writing Section*

Task 1 Item ID	Task 1 Round 1 mean (SD)	Task 1 Round 2 mean (SD)	Task 2 Item ID	Task 2 Round 1 mean (SD)	Task 2 Round 2 mean (SD)
TSK1-01	1.81 (.40)	1.94 (.25)	**TSK2-01**	1.63 (.50)	1.69 (.48)
TSK1-02	1.75 (.45)	1.94 (.25)	**TSK2-02**	2.00 (.52)	2.00 (.00)
TSK1-03	-	-	**TSK2-03**	-	-
TSK1-04	1.44 (.51)	1.06 (.25)	**TSK2-04**	1.13 (.34)	1.00 (.00)
TSK1-05	2.75 (.45)	2.94 (.25)	**TSK2-05**	1.94 (.25)	2.00 (.00)
TSK1-06	2.13 (.34)	2.06 (.25)	**TSK2-06**	2.50 (.52)	2.63 (.50)
TSK1-07	2.06 (.25)	2.00 (.00)	**TSK2-07**	2.00 (.52)	2.13 (.34)
TSK1-08	1.06 (.25)	1.06 (.25)	**TSK2-08**	1.31 (.48)	1.19 (.40)
TSK1-09	1.00 (.00)	1.00 (.00)	**TSK2-09**	-	-
TSK1-10	1.63 (.50)	1.56 (.51)	**TSK2-10**	1.50 (.52)	1.44 (.51)
TSK1-11	1.31 (.48)	1.31 (.48)	**TSK2-11**	1.13 (.34)	1.06 (.25)
TSK1-12	1.94 (.25)	2.00 (.00)	**TSK2-12**	1.75 (.45)	1.81 (.40)
TSK1-13	2.69 (.48)	2.88 (.34)	**TSK2-13**	1.31 (.48)	1.31 (.48)
TSK1-14	1.00 (.00)	1.00 (.00)	**TSK2-14**	-	-
TSK1-15	1.44 (.63)	1.44 (.51)	**TSK2-15**	2.50 (.63)	2.56 (.51)
TSK1-16	2.19 (.75)	2.31 (.60)	**TSK2-16**	2.44 (.51)	2.69 (.48)
TSK1-17	2.81 (.40)	2.94 (.25)	**TSK2-17**	2.19 (.66)	2.25 (.58)
TSK1-18	2.75 (.45)	2.81 (.40)	**TSK2-18**	2.81 (.40)	2.81 (.40)
TSK1-19	1.50 (.63)	1.38 (.62)	**TSK2-19**	3.00 (.00)	3.00 (.00)
TSK1-20	2.00 (.37)	2.06 (.73)	**TSK2-20**	2.56 (.51)	2.69 (.48)
TSK1-21	2.00 (.73)	2.00 (.73)	**TSK2-21**	1.94 (.25)	1.88 (.34)
TSK1-22	2.75 (.45)	2.81 (.40)	**TSK2-22**	1.00 (.00)	1.13 (.50)
TSK1-23	2.44 (.51)	2.38 (.50)	**TSK2-23**	2.75 (.45)	2.81 (.40)
TSK1-24	1.88 (.34)	1.88 (.34)	**TSK2-24**	-	-
TSK1-25	2.63 (.50)	2.63 (.50)	**TSK2-25**	2.31 (.48)	2.25 (.45)

(*Note: Dashes indicate that the scripts were not accessed)*

An examination of the data reveals that the standard deviation (indicated in parenthesis) of nearly all of the scripts decreased between rounds, suggesting that the discussion that took place between rounds helped judges come to a consensus on the level of each script. Some of the scripts (TSK1-03, TSK2-03, TSK2-09, TSK2-14, TSK2-24) were not assessed as the quality of the photocopies made them difficult to read and judges were instructed to ignore those scripts.

The judges' mean scores were rounded to the nearest integer and then cross tabulated with the official benchmarked scores for each script. Tables 4.6 and 4.7 illustrate Round 1 and Round 2 judge mean CEFR scores cross tabulated against the official benchmarked scores. The top row shows the official benchmarked score each script received and the first column shows the CEFR level the judges matched each script to. In both rounds, the minimum score that was assigned a B1 level was 11 (shaded cell).

***Table 4.6** Round 1 Cut Score Judgment for the Writing Section*

Score	4	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total
A2	4	1	2	1	3	0	0	0	0	0	0	0	0	0	0	11
B1	0	0	1	3	1	3	4	1	3	1	1	2	1	1	0	22
B2	0	0	0	0	0	0	0	0	0	2	0	2	2	2	4	12
Total	4	1	3	4	4	3	4	1	3	3	1	4	3	3	4	45

***Table 4.7** Round 2 Cut Score Judgment for the Writing Section*

Score	4	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total
A2	4	1	2	3	3	0	0	0	0	0	0	0	0	0	0	13
B1	0	0	1	1	1	3	4	1	3	1	1	1	1	1	0	19
B2	0	0	0	0	0	0	0	0	0	2	0	3	2	2	4	13
Total	4	1	3	4	4	3	4	1	3	3	1	4	3	3	4	45

4.4 Judgments of GVR Section

Table 4.8 illustrates Round 1 and Round 2 mean scores for the GVR section. In both rounds, the cut score mean was 40 (shaded cells) when rounded to the nearest integer. The standard deviation (SD) between rounds increased from 7.88 to 8.35. This increase in variance was attributed to judges changing some of their estimates based on the *reality information* (p-values) received before Round 2. The inter-rater reliability index (ICC) increased from .79 in Round 1 to .83 in Round 2.

***Table 4.8** Cut Score Judgments for the GVR Section*

Judge ID	Round 1	Round 2
J1	30.20	31.45
J2	46.15	54.15
J3	37.98	37.62
J4	44.92	45.29
J5	30.72	25.67
J6	33.48	33.26
J7	38.76	39.26
J8	44.56	42.20
J9	50.90	48.28
J10	56.63	54.22
J11	46.90	43.30
J12	37.10	39.07
J14	34.95	37.10
J15	30.98	29.87
J17	40.90	44.51
Mean	40.34	40.35
SD	7.88	8.35
Min	30.20	25.67
Max	56.63	54.22
Sec	2.04	2.15
ICC	0.79	0.83
Number of Judges	**15**	**15**

(Note: J13 could not attend this session).

4.5 Judgments of Listening Section

Table 4.9 illustrates Round 1 and Round 2 mean scores for the Listening section. In both rounds, the cut score (mean) was 18 (shaded cells) when rounded to the nearest integer. The standard deviation (SD) between rounds increased from 2.80 to 2.98. Once again this increase in variance may have been attributed to judges changing some of their estimates based on the *reality information* (p-values) received before Round 2. The inter-rater reliability (ICC) index increased from .71 in Round 1 to .74 in Round 2.

***Table 4.9** Cut Score Judgments for the Listening Section*

Judge ID	Round 1	Round 2
J1	18.60	16.00
J2	23.90	23.45
J3	19.23	18.42
J4	17.54	19.59
J5	17.95	15.01
J6	13.32	13.32
J7	16.30	16.15
J8	19.20	19.20
J9	17.93	17.73
J10	15.59	15.82
J11	18.80	18.90
J12	15.75	16.64
J14	19.90	19.95
J15	13.06	11.95
J17	20.74	20.74
Mean	17.85	17.52
SD	2.80	2.98
Min	13.06	11.95
Max	23.90	23.45
Sec	0.72	0.77
ICC	0.71	0.74
Number of Judges	**15**	**15**

(Note: J13 could not attend this session).

4.6 Cut Score Validation Analyses

The rest of this section deals with the cut score validation analyses that were run in order to examine consistency within the method, intraparticipant consistency, interparticipant consistency, and decision consistency.

The initial recommended cut scores rounded to the nearest integer based on the Round 2 estimates were the following:

Speaking section: 12
Writing section: 11
Listening section: 18
GVR section: 40

However, prior to presenting any validation analyses, the psychometric characteristics of the BCCE™ examination on which the cut scores for the GVR and Listening sections are determined will be presented. The assumption behind any standard setting workshop is that the examination on which the cut scores are to be determined should exhibit good psychometric characteristics; otherwise, the recommended cut scores may be invalid (Council of Europe, 2009). Table 4.10 illustrates the psychometric characteristics of the GVR and Listening sections.

***Table 4.10** Psychometric Characteristics of the GVR and Listening Sections*

N = 921*	GVR section	Listening section
Number of items	75	30
Mean	46.75	20.43
SD	14.32	5.32
Min Score	12	5
Max Score	74	30
Mean P	.62	.68
Mean R-pbis	.38	.32
Alpha	.93	.81
SEM	3.68	2.30

Note: ** Data available at the time of the study.*

The reliability (alpha) of the Grammar and Listening sections is .93 and .81 respectively, implying that both sections exhibit acceptable reliability estimates and that valid cut scores can be determined for these sections.

4.6.1 Consistency within the Method

Hambleton and Pitoniak (2006) define consistency within the method as "the extent to which same performance standards would be obtained if the method were replicated" (p.458). The first analysis examined method consistency by estimating the standard error of the cut score (SEc). The equation used to calculate the standard error of the cut score is the following:

$$SEc = \frac{SDs}{\sqrt{n}}$$

In the above equation, the standard error of the cut score (SEc) is equal to the standard deviation of the individual cut scores (SDs) divided by the square root of the number of judges (CEFR Manual, 2009: 104). The SEc "is one of the classical indices indicative of the replicability of the obtained results" (Kaftandjieva, 2010: 103) and is compared to the standard error of measurement (SEM) of the test.

Several criteria have been suggested when comparing the SEc to the SEM. Jaeger (1991) suggests that the SEc should be no greater than one quarter of the SEM, while Cohen, Kane, and Crooks (1999) state that the SEc should not be greater than half the SEM. On the other hand, Kaftandjieva (2010) recommends a compromise between the previous two criteria and suggests that the SEc should be no greater than a third of the SEM.

The criterion used for this study was Cohen's et.al (1999) who claim that a SEc that is equal to or smaller than one half the SEM "will have relatively little impact on the misclassification rates" (p.364). Applying Cohen's et al. (1999) criterion as an internal check would mean that the SEc divided by the SEM should be equal to or less than .5 (SEc/SEM ≤ .50).

Table 4.11 displays the standard error of the cut score and the standard error of measurement for both the GVR and Listening sections. The first row displays the section and the round. The second row illustrates the judges' standard error of cut score (SEc) based on Round 2 estimates. The third row illustrates the test's standard error of measurement (SEM). The fourth row illustrates the internal check estimate (SEc/SEM).

Table 4.11 Standard Errors of Cut Scores and Measurement

	GVR (Round 2)	**Listening (Round 2)**
SEc	2.04	.77
SEM	3.68	2.30
SEc/SEM	.55	.33

The GVR and Listening sections had a SEM of 3.68 and 2.30 and a SEc of 2.04 and .77 respectively. However, for there to be little impact on classification rates, the SEc of the GVR section would have to be equal to or less than 1.84 $\left(\frac{3.68}{2}\right)$ and the SEc of the Listening section would have to be equal to or less than 1.15 $\left(\frac{2.30}{2}\right)$.

This criterion was only met for the Listening section since the Listening section SEc (.77) was less than half the Listening section SEM (2.30). Thus, in order for the Grammar section SEc (2.04) to be minimized, the judges' GVR section Round 2 mean was trimmed (Zieky, Perie, & Livingston, 2008) by excluding extreme judge ratings - ratings too high or too low - until the SEc was equal to or below 1.84. It should be noted that despite the fact the SEc was acceptable for the Listening section, the Listening section Round 2 mean was also trimmed as the inter-rater reliability ICC for the Listening section was below the minimum acceptable level (0.8).

Table 4.12 illustrates the effect the trimming of the judges' mean had on both the mean cut score and the inter-rater reliability of the judges. For both trimmed rounds, the two highest and the two lowest ratings were excluded. The SEc for the GVR section decreased from 2.15 to 1.56 and from .77 to .52 in the Listening section. The standard deviation decreased from 8.35 to 5.16 in the GVR section and from 2.98 to 1.74 in the Listening section. The inter-rater reliability (ICC) increased from .83 to .86 in the GVR section and from .74 to .84 in the Listening section. However, in both cases, the cut scores for Round 2 with extreme ratings and without extreme ratings were identical when rounded to the nearest integer (see shaded cells).

***Table 4.12** Recommended Cut Score for the Listening and GVR Sections*

Judge ID	GVR Round 2 (with extreme ratings)	GVR Round 2 (without extreme ratings)	Listening Round 2 (with extreme ratings)	Listening Round 2 (without extreme ratings)
J1	31.45	31.45	16.00	16
J2	54.15	(excluded)	23.45	(excluded)
J3	37.62	37.62	18.42	18.42
J4	45.29	45.29	19.59	19.59
J5	25.67	(excluded)	15.01	(excluded)
J6	33.26	33.26	13.32	(excluded)
J7	39.26	39.26	16.15	16.15
J8	42.20	42.20	19.20	19.2
J9	48.28	48.28	17.73	17.73
J10	54.22	(excluded)	15.82	15.82
J11	43.30	43.30	18.90	18.9
J12	39.07	39.07	16.64	16.64
J14	37.10	37.10	19.95	19.95
J15	29.87	(excluded)	11.95	(excluded)
J17	44.51	44.51	20.74	20.74
Mean	40.35	40.12	17.52	18.10
SD	8.35	5.16	2.98	1.74
Min	25.67	31.45	11.95	15.82
Max	54.22	48.28	23.45	20.74
SEc	2.15	1.56	0.77	0.52
ICC	0.83	0.86	0.74	0.84
Number of Judges	**15**	**11**	**15**	**11**

4.6.2 Intraparticipant Consistency

Hambleton and Pitoniak (2006) define intraparticipant consistency as "the degree to which a panelist is able to provide ratings that are consistent with the empirical difficulties and the degree to which ratings change across rounds" (p.458). Consistency between empirical difficulties and judge estimates also provides further support for the adequacy of a standard-setting method (Chang, 1999) and "is one of the main criteria for internal validity and the adequacy of the cut scores" (Kaftandjieva, 2010: 50).

Table 4.13 shows the intraparticipant consistency for all four BCCE™ examination sections. For the Speaking and Writing sections, the judges' item mean score was correlated with the benchmarked scores. For the Listening and GVR sections, the judges' item mean was correlated with the empirical difficulty of the item. According to Brandon (2004), the means of the judges' estimates should "correlate well with the actual difficulty levels for the items" (p.71) so that the estimates can be considered valid. Hambleton et. al (2012) claim that low correlation estimates may indicate that the judges' content knowledge needs to be questioned.

***Table 4.13** Intraparticipant Consistency: Judgments with Empirical P-values*

	Round 1	Round 2	Round 2 Trimmed Mean
Speaking	.95	.94	
Writing	.88	.87	
GVR	.38	.81	.75
Listening	.68	.90	.92

(*Note: all correlations were significant at p ≤ .01).*

For the Speaking section, the Spearman correlation between the judges' item mean score and the benchmarked scores decreased from .95 for Round 1 to .94 for Round 2. For the Writing section, the spearman correlation between the judges' item mean score and the benchmarked scores decreased from .88 for Round 1 to .87 for Round 2. For the GVR section, the Spearman correlation between the judge's item mean estimate and the empirical item mean estimate dramatically increased from .38 for Round 1 to .81 for Round 2 and then decreased to .75 for Round 2 Trimmed Mean. For the Listening section, the Spearman correlation between the judge's item mean estimate and the empirical item mean estimate increased from .68 for Round 1 to .90 for Round 2 to .92 for Round 2 trimmed mean. *The high correlations observed in all four sections at the end of Round 2 add evidence to intraparticipant consistency and to the panels' content expertise.*

Table 4.14 illustrates the degree to which the ratings of each judge changed across rounds for the GVR and Listening sections. Hambleton, Pitoniak, and Copella (2012) claim that when judges do not change their ratings between rounds, they may not be considering feedback provided between rounds.

***Table 4.14** Intraparticipant Consistency: Changes in Ratings across Rounds*

Judge ID	GVR Round 1 and Round 2 Difference	Listening Round 1 and Round 2 Difference
J1	-1.25	2.6
J2	-8.00	0.45
J3	0.36	0.81
J4	-0.37	-2.05
J5	5.05	2.94
J6	0.22	0
J7	-0.50	0.15
J8	2.36	0
J9	2.62	0.20
J10	2.41	-0.23
J11	3.60	-0.10
J12	-1.97	-0.89
J14	-2.15	-0.05
J15	1.11	1.11
J17	-3.61	0

For the GVR section, all judges changed their ratings across rounds. The difference between Round 1 and Round 2 ranged from -8.00 for J2 to 5.05 for J5. For the Listening section, not all of the judges changed their ratings across rounds. The difference between Round 1 and Round 2 ranged from -2.05 for J4 to 2.6 for J1 (see Tables 4.8 and 4.9 for actual estimates). Three judges (J6, J8, & J17) entered the same ratings for both rounds. The fact that these three judges had no change in their Listening section ratings between rounds may not necessarily imply that the judges did not consider the reality information presented or what was discussed during the Round 1 discussion. The three judges may have felt comfortable with their original

ratings and did not find it necessary to change them. *Nonetheless, the changes in ratings across rounds suggests that the majority of the judges considered the feedback presented between rounds, adding further evidence of intraparticipant consistency.*

4.6.3 Interparticipant Consistency

Hambleton and Pitoniak (2006) define interparticipant consistency as "the consistency of item ratings and performance standards across panelists" (p.458). Table 4.15 displays the inter-rater agreement indices for all sections across all rounds.

***Table 4.15** Interparticipant Agreement and Consistency*

Section	RAI	Cronbach's alpha	ICC
Speaking Round 1	.88	.98	.98
Speaking Round 2	.90	.99	.99
Writing Round 1	.82	.97	.96
Writing Round 2	.86	.98	.98
GVR Round 1	.85	.86	.79
GVR Round 2	.87	.90	.83
GVR Round 2 (trimmed mean)	.90	.90	.86
Listening Round 1	.88	.79	.72
Listening Round 2	.88	.83	.74
Listening Round 2 (trimmed mean)	.89	.88	.84

For the Speaking section, the RAI increased from .88 for Round 1 to .90 for Round 2. Cronbach's alpha increased from .98 for Round 1 to .99 for Round 2 and the ICC increased from .98 for Round 1 to .99 for Round 2.

For the Writing section, the RAI increased from .82 for Round 1 to .86 for Round 2. Cronbach's alpha increased from .97 for Round 1 to .98 for Round 2 and the ICC increased from .96 for Round 1 to .98 for Round 2.

For the GVR section, the RAI increased from .85 for Round 1 to .87 for Round 2. Cronbach's alpha increased from .86 for Round 1 to .90 for Round 2 and the ICC increased from .79 for Round 1 to .83 for Round 2. Even though the changes in RAI and ICC were small, they were consistent. However, the GVR Round 2 estimates were trimmed and this resulted in the RAI increasing from .87 for Round 2 to .90 for the trimmed round. Cronbach's alpha remained the same between Round 2 and the trimmed round, but the ICC increased from .83 for Round 2 to .86 for the trimmed round. *The increases in the RAI and the ICC provide further evidence that trimming had a positive effect on section cut score confidence and the inter-rater reliability of the judges.*

For the Listening section, the RAI remained the same between Round 1 and Round 2. Cronbach's alpha increased from .79 for Round 1 to .83 for Round 2 and the ICC increased from .72 for Round 1 to .74 for Round 2. However, the Listening Round 2 estimates were trimmed and this resulted in the RAI increasing from .88 for Round 2 to .89 for the trimmed round. Cronbach's alpha increased from .83 for Round 2 to .88 for the trimmed round and the ICC increased from .74 for Round 2 to .84 for the trimmed round. Similarly, to the GVR section, trimming Round 2 estimates had a positive effect on section cut score confidence and the inter-rater reliability of the judges.

*The high inter-rater reliability for Rounds 2 and the trimmed rounds indicate that the judges were "very homogeneous in terms of exact agreement as well as in terms of association" (*Kaftandjieva & Takala, 2002: 113).

4.7 Decision Consistency and Accuracy Analyses

"Decision consistency refers to the agreement between the classifications of the same candidates on two different examinations with the same test" (Kaftandjieva, 2004:26). However, in order to compute such coefficients candidates would have to take the same examination twice.

Some methods (Livingston & Lewis 1995; Subkoviak, 1988) have been proposed in the literature to estimate decision consistency and accuracy based on a single administration. These methods provide the "likelihood that an examinee classified as passing (or failing) on one administration of an examination will be classified similarly on a second administration" (Cizek & Bunch, 2007: 309).

The Subkoviak method

The Subkoviak method for examining decision consistency based on a single administration was applied using the following equation:

$$Z = \left(\frac{Cx - M - 0.5}{Sx}\right)$$

In this equation, Z is the standard score, C_x is the cut score for the test, M is the observed test mean and S_x is the standard deviation (SD) of observed test scores. Absolute values of Z (IZI) are then used to obtain the estimates of agreement coefficient (p_0) and kappa (k) from two tables using the reliability estimate for the total test scores.

For example, in the GVR section, the recommended cut score (C_x) was 40, the observed test mean (M) was 46.75, and the standard deviation (SD) of observed test scores was 14.32. Applying these values in the equation above Z is as follows:

$$Z = \left(\frac{40 - 46.75 - 0.5}{14.32}\right) = -0.51$$

and $IZI = I\text{-}0.51I = 0.51$

The test reliability (alpha) of the GVR section of the BCCE™ examination was .93. Thus, using Subkoviak's table for approximate values of the agreement coefficient (p_0) for various values of reliability (Subkoviak, 1988) a IZI value of .51 and a test reliability of .93 yields an agreement coefficient (p_0) of .87, implying that 87% of the candidates would be consistently classified as masters or non masters had two equivalent test administrations been employed. Using Subkoviak's table for approximate values of kappa for various values of reliability (Subkoviak, 1988), a IZI value of .51 and a test reliability of .93 yields a corrected decision consistency agreement coefficient, (k), of .70. The kappa coefficient is used to indicate the consistency of the classifications beyond what would be expected by chance. The maximum value for p_0 is .98 and .71 for k. Subkoviak (1988) suggests that a test should be long enough so that the kappa coefficient is "within the approximate range .60 and .70" (p.53). Table 4.16 illustrates the Agreement coefficient (p_0) and kappa (k) estimates for the Listening and GVR sections.

***Table 4.16** Agreement Coefficient (p_0) and kappa (k) for the GVR and Listening Sections*

	p_0	k
GVR section	.87	.70
Listening section	.80	.59

The agreement coefficient (p_0) for the GVR and Listening sections were .87 and .80 respectively, and the kappa coefficient (k) for the GVR and Listening sections were .70 and .59 respectively. The moderate kappa value for the Listening section may have been attributed to the fact that the Listening section contained only 30 items and/or that the reliability of the Listening section was below .90. In order to obtain a kappa coefficient to be at least .60 when using Subkoviak's table for approximate values of kappa, the reliability of a test must be at least .90.

Applying Subkoviak's (1988) 'rule of thumb' for kappa coefficient values, both kappa values are very close to or within the approximate range of .60 and .70.

The Livingston and Lewis method:

The Livingston and Lewis method is "a generally applicable method for using data from one form of a test to estimate the accuracy and the consistency of classifications based on the scores" (Livingston & Lewis, 1995: 179). The Livingston and Lewis decision consistency and accuracy estimates were obtained by using the BB – Class software (Brennan, 2001). The four-parameter beta binomial model was selected for analysis. The method produces test taker classification estimates that "tend to be within one percentage point of their actual values" (Livingston & Lewis, 1995: 196).

Table 4.17 illustrates decision accuracy by comparing observed score classifications with estimated true score classifications. The first row shows the probability of correct classification, the next two rows show the probability of false positive and false negative errors. False positive errors occur when candidates are estimated to be above the cut score when in fact they are not. Similarly, false negative errors occur when candidates are estimated to be below the cut score, when in fact they are not (Hambleton & Novick, 1973).

***Table 4.17** Accuracy Relative to Actual Observed Scores*

	Speaking	Writing	GVR	Listening
Probability of correct classification	.92	.95	.93	.89
False positive	.05	.02	.03	.05
False negative	.03	.03	.04	.06

The probability of correct classification of candidates for all four sections of the BCCE™ examination ranges from .89 for the Listening section to .95 for the Writing section. The probability of false positive classifications range from .02 for the Writing section to .05 for the Speaking and Listening sections and the probability of false negative classifications range from .03 for the Speaking and Writing sections to .06 for the Listening section.

Table 4.18 illustrates decision consistency by comparing classification decisions made based on expected and observed scores. The first row illustrates the overall percentage of consistent classifications, the second row indicates the probability of misclassifications.

***Table 4.18** Consistency Using Expected (Row) vs. Actual (Column) Observed Scores*

	Speaking	Writing	GVR	Listening
Overall percentage of consistent classifications (p_c)	.88	.94	.90	.85
kappa (k)	.71	.79	.78	.64
Probability of misclassification	.12	.06	.10	.15

The overall percentage of consistent classifications (p_c) of candidates for all four sections of the BCCE™ examination ranges from .88 for the Speaking section to .94 for the Writing section. The kappa coefficient value (k) ranges from .64 for the Listening section to .79 for the Writing section. *Applying Subkoviak's (1988) 'rule of thumb' for kappa coefficient values, all kappa values are within the approximate range of .60 and .70 or higher.* The probability of misclassifications ranges from .06 for the Writing section to .15 for the Listening section.

Table 4.19 illustrates the judges' inter-rater reliability indices for Round 2. For the GVR and Listening sections, the inter-rater reliability indices are calculated on the trimmed mean (without the extreme judgments).

***Table 4.19** Benchmarking and Standard Setting Inter-rater Reliability Indices*

	Speaking	Writing	GVR	Listening
Rater agreement index (RAI)	.90	.86	.90	.89
Cronbach's alpha	.99	.98	.86	.88
ICC	.99	.98	.90	.84

The RAI for each section ranged from .86 for the Writing section to .90 for the Listening and GVR sections. The Cronbach's alpha estimate ranged from .86 for the GVR section to .99 for the Speaking section and the ICC estimate ranged from .84 for the Listening section to .99 for the Speaking section. *The inter-rater reliability indices for all sections were high. It can, therefore, be concluded that the BCCE™ recommended examination section cut scores for CEFR level B1 fulfill generally accepted criteria for decision accuracy, consistency, and agreement.*

The final *recommended* cut scores were the following:

Speaking section: 12
Writing section: 11
GVR section: 40
Listening section: 18

4.8 Final Cut Score Establishment

In August 2011, the policy committee convened to evaluate the recommended cut scores. The committee members carefully examined the data presented in Section 4 of this report, Tables 4.20 to 4.23 illustrating decision accuracy and consistency estimates of the recommended section cut scores +/- one raw score point, and the overall pass rate on each section (see Hellenic American University, BCCE™ 2011 Test Administration Report for final pass rate analyses). A brief summary of the committee's decision follows each table.

***Table 4.20** Speaking Section Cut Score Evaluation*

	11	12 (recommended)	13
Probability of correct classification	.92	.92	.91
False positive rate	.04	.05	.05
False negative rate	.03	.03	.04
Probability of consistent classification	.89	.88	.87
kappa	.69	.71	.74
Probability of misclassifications	.11	.12	.13

Considering that the interparticipant consistency estimates (see Table 4.15) for the Speaking section were high and that the probability of correct classification was the same for a score of 11 and 12, the committee accepted the recommended cut score of 12 as it resulted in a higher kappa.

***Table 4.21** Writing Section Cut Score Evaluation*

	10	11 (recommended)	12
Probability of correct classification	.97	.96	.95
False positive rate	.02	.02	.03
False negative rate	.01	.02	.02
Probability of consistent classification	.95	.94	.93
kappa	.77	.79	.81
Probability of misclassifications	.05	.06	.07

Considering that the interparticipant consistency estimates (see Table 4.15) for the Writing section were high and that the probability of correct classification was almost the same for a score of 10 and 11, the committee accepted the recommended cut score of 12 as it resulted in a higher kappa.

***Table 4.22** GVR Section Cut Score Evaluation*

	39	40 (recommended)	41
Probability of correct classification	.93	.88	.91
False positive rate	.03	.11	.08
False negative rate	.04	.01	.01
Probability of consistent classification	.90	.90	.90
kappa	.78	.78	.78
Probability of misclassifications	.10	.10	.10

Considering the interparticipant consistency estimates (see Table 4.15) and that the recommended cut score was trimmed for the GVR section, the committee lowered the cut score to 39. A score of 39 displayed the highest probability of correct and consistent classifications and resulted in the lowest probability of false positive and negative rates in relation to the two other scores above it.

***Table 4.23** Listening Section Cut Score Evaluation*

	17	18 (recommended)	19
Probability of correct classification	.90	.89	.88
False positive rate	.05	.05	.06
False negative rate	.05	.06	.06
Probability of consistent classification	.86	.85	.84
kappa	.62	.64	.65
Probability of misclassifications	.14	.15	.16

Considering the interparticipant consistency estimates (see Table 4.15) and that the recommended cut score was trimmed for the Listening section, the committee lowered the cut score to 17. A score of 17 displayed the highest probability of correct and consistent classifications and resulted in the lowest probability of misclassifications and false positive and negative rates in relation to the other two scores above 17.

Table 4.24 illustrates the final cut scores for the BCCE™ examination. The first row illustrates the section and the second row illustrates the final raw cut score. The third row shows the Rasch equivalent cut score.

***Table 4.24** Final Cut Scores*

Section	Raw Score	Ability estimate (Rasch)
Speaking	12	
Writing	11	
GVR	39	.13
Listening	17	.29

The four sections of the BCCE™ examination are scored using an advanced mathematical model (Rasch). For each section of the BCCE™ examination, ability estimates are converted into scaled scores ranging from 0 to 30. For the Listening and GVR sections, scaled scores are not simply the percentage of correct answers because the ability of candidates and the difficulty of an item have been incorporated in the scores. Similarly, for the Speaking and Writing sections, the difficulty of the task as well as the leniency and/or severity of the rater have been incorporated in the scaled scores.

For the GVR section a raw score of 39 was equivalent to an ability estimate of .13 (Rasch) and for the Listening section a raw score of 17 was equivalent to an ability estimate of .29 (Rasch). All BCCE™ examinations are equated so that the difficulty of the test remains constant from form to form and year to year. Therefore, the cut score ability estimates for the GVR and Listening sections are .29 and .13 respectively.

The BCCE™ examination results for each section are reported on a scale of 0 – 30. The minimum scaled pass score for each section and the examination are shown in Table 4.25.

***Table 4.25** Section and Total Scaled Scores*

Section	Minimum scaled pass score
Listening	19/30
GVR	19/30
Writing	18/30
Speaking	18/30
Total Scaled Score	**74/120**

To pass the Listening and Grammar sections, candidates need to achieve a scaled score of at least 19 out of 30, while to pass the Writing and Speaking sections candidates need to get a scaled score of at least 18 out of 30. The CEFR acknowledges that some language learners may exhibit uneven profiles (Council of Europe, 2001) since they "differ in their profile of skills" (Council of Europe, 2011: 13). For example, some learners are better at receptive skills than productive skills. Thus, a compensatory standard setting strategy (Hambleton & Pitoniak, 2006; Kaftandjieva, 2010) is used in calculating the final BCCE™ examination pass/fail grade. Candidates who "perform below a typical passing level" (Haladyna & Hess, 1999) on one section *only*, still receive an overall pass when the sum of the scaled scores for each section is at least 74 out of a total of 120. The assumption behind using this strategy is that "the sum of the separate components reflects adequately the measured construct" (Kaftandjieva, 2010: 15). Haladyna and Hess (1999) claim in a compensatory standard setting strategy "the reliability of the total scores tends to be very high" (p.134).

4.9 External Validation

Hambleton and Pitoniak (2006) suggest three ways that external validation evidence can be collected: (1) comparisons to other standard setting methods; (2) comparisons to other sources; and (3) reasonableness of performance levels.

Only one standard setting method was selected for each section of the BCCE™ examination. A second method was not selected as it would entail additional workshop days and would add more of a cognitive demand on the judges. Following Cizek and Bunch (2007), the standard setting and benchmarking methods were selected for their "strong match with the format of the assessment, the purposes of testing, the skills of the participants, and other factors" (p.39).

An attempt was made to collect data from the candidates by asking them whether they were taking or had taken another B1 test around the same time they took the BCCE™ examination; however, the candidates were reluctant to offer such information. Another attempt was made to obtain information from another examining board so that the results of candidates that had taken two B1 examinations around the same time could be compared. However, the examining boards' local representatives were reluctant to exchange information on candidates sitting both examinations.

As the BCCE™ examination is continually increasing in terms of test population size and geographical locations where it is being administered, another standard setting cut score study will take place. External validation will be presented in that study as data from candidates and their teachers will be collected and correlated with actual BCCE™ candidate examination scores.

4.10 The Judges' Feedback

On the last day of the workshop, judges anonymously completed an evaluation of the four-day workshop. Judges were asked to indicate their satisfaction with their training, and to indicate their level of confidence in the results and the defensibleness and reasonableness of the recommended final cut score.

Fifteen questionnaires were administered, with quantitative data in the form of a four-point Likert scale and qualitative data in the form of free responses. The judges' responses are summarized in Table 4.26. For the quantitative section of the evaluation form, judges were asked to choose a rating for each statement according to the following criteria: 4 = strongly agree; 3 = agree; 2 = disagree; and 1 = strongly disagree.
All of the judges reported that they were confident about the defensibility and appropriateness of their final recommended cut scores and that the familiarization activities completed and the training they received helped them understand how to perform each task and to complete their ratings accurately. Nearly all of the judges (93%) reported that the standard setting process was clearly explained.

When judges reported on what they liked *best* about the workshop, comments ranged from the pleasant and friendly environment created to the training and feedback received. When judges reported on what they liked *least* about the workshop, comments ranged from the amount of noise made by some other judges during familiarization tasks, which was immediately addressed at the time, to the difficulty of carrying out online tasks as no underlining could take place.

Overall, the quantitative ratings and qualitative feedback received were favorable adding further evidence of procedural validity (Kane, 1994).

***Table 4.26** Judges' Evaluation Form Responses*

Statements	(4)	(3)	(2)	(1)	Missing	Total	Mean (/ 4)
1. The PowerPoint Presentation on Day 1 provided me with a clear understanding of the purpose of a CEFR Linking Project.	8	5	1	0	1	15	3.5
2. The coordinator clearly explained the Standard Setting process.	10	4	1	0	0	15	3.6
3. The familiarization activities were helpful in understanding CEFR levels.	13	2	0	0	0	15	3.87
4. The amount of familiarization activities was enough to achieve a better understanding of the CEFR levels.	11	3	0	0	1	15	3.79
5. The training activities were helpful in understanding CEFR levels.	13	2	0	0	0	15	3.87
6. The training activities were enough to achieve a better understanding of the CEFR levels.	9	6	0	0	0	15	3.60
7. The training and activities helped me understand how to perform each task.	12	3	0	0	0	15	3.80
8. The panel discussions and small group discussions were helpful in understanding the process.	8	6	0	0	1	15	3.57
9. The time spent on the discussions was enough.	10	4	0	0	1	15	3.71
10. I felt that I had an equal opportunity to contribute ideas and opinions in the panel discussion.	13	2	0	0	0	15	3.87
11. I felt that I had an equal opportunity to contribute ideas and opinions in the small group discussions.	12	3	0	0	0	15	3.80
12. The discussions after the first round of my ratings were helpful to me.	8	6	0	0	1	15	3.57
13. The feedback I received on my ratings was helpful to me.	7	7	0	0	1	15	3.50
14. The student statistics presented were easy to follow.	5	9	1	0	0	15	3.27
15. The student statistics presented helped me to finalize my ratings.	7	6	2	0	0	15	3.33
16. I was able to follow the instructions and complete the rating accurately.	11	4	0	0	0	15	3.73
17. The discussions after the second round of ratings were helpful to me.	9	6	0	0	0	15	3.60
18. I am confident about the defensibility and appropriateness of the final recommended cut scores.	7	8	0	0	0	15	3.47
19. The facilities and food service helped create a productive and efficient working environment.	9	6	0	0	0	15	3.60
20. I felt comfortable using a PC on Day 1.	7	5	3	0	0	15	3.27
21. I felt comfortable using a PC on Day 4.	12	3	0	0	0	15	3.80
22. I enjoyed using a PC.	13	1	0	0	1	15	3.93
23. The coordinator guided the sessions effectively.	12	2	0	0	1	15	3.86
				Overall Average			**3.64**

(Note: J13 was absent on the last day).

Question: **What did you like *best* about the sessions?**

Comments:

- The coordinator perfectly combined professional work with a relaxed atmosphere and humor.
- The explanations of the student statistics.
- Taking the test as a candidate helped me understand better or remember how the learner handles each item.
- Pleasant environment; pleasant coordinator; pleasant fellow workshop attendees.
- Friendly environment; good speaking to/discussing with other attendees.
- Coordinator kept us 'awake'. Stepped in when necessary. Repeated instructions. Allowed us time to think, reflect and discuss.
- Good experience!
- The way things were explained clearly and precisely. Our trainer was very patient and informative.
- The training activities and the small group discussions.
- The presentations and feedback.

Question: **What did you like *least* about the sessions?**

Comments:

- Too much noise from people talking when they were not supposed to.
- Taking the CEFR familiarization activities on the computer (the ones we had to match sentences to levels). I generally need to underline and even compare to do such activities and the PC screen didn't help me.
- Too much paperwork; more PC use would be beneficial next time.
- During the speaking, more videos than necessary were shown.
- When I could not understand how scales were applied (few times); when I felt I had to go against a previous evaluation; but that was just personal judgment call.

Adapted from Cizek & Bunch (2007)

5. Conclusion

This technical report has presented the setting of CEFR level B1 cut scores for the four sections of the BCCE™ examination.

A panel of 16 experts participated in the standard setting workshop. The Benchmarked method and the Modified Angoff method were applied to the constructed-response questions and the selected-response questions respectively. During most of the workshop, judges used computers to enter their ratings directly into a Microsoft Excel® worksheet. By using this medium, data processing was sped up (Hambleton, Pitoniak, & Copella, 2012) and judges received feedback on their rankings and ratings in a very short time.

One the basis of evidence collected and the results of detailed statistical analyses documented here, it is concluded that the cut scores are well aligned to the CEFR B1 level.

References

Assessment Systems Corporation. (2009). Angoff Analysis Tool [computer software]. Available at http://www.assess.com (Last accessed July 25, 2011)

Blackboard Inc. (2010). Blackboard Learn™ version 9.1 [online learning platform]. Available at http://www.blackboard.com

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education*, *17* (1), 59 – 88.

Brennan, R. L. (2001). BB-CLASS (version 1.1) [Computer software]. Iowa City: Center for Advanced Studies in Measurement, University of Iowa.

Brennan, R. L. (2004). Manual for BB-Class: A computer program that uses the Beta-Binomial model for classification consistency and accuracy [Computer manual]. *Casma Research Report (9)*, 1 - 22.

Burry-Stock, J.A., Shaw, D.G., Laurie, C., & Chissom, B.S. (1996). Rater agreement indexes for performance assessment. *Educational and Psychological Measurement, 56* (2), 251 – 262.

Chang L. (1999). Judgmental item analysis of the Nedelsky and Angoff standard setting methods. Applied measurement in Education, 12 (2), 151 – 165.

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

Council of Europe. (2005*). Relating language examinations to the Common European Framework of Reference for languages: learning, teaching, assessment (CEFR): Reading and Listening Items and Tasks: Pilot Samples illustrating the common reference levels in English, French. German, Italian and Spanish* [CD]. Strasbourg: Language Policy Division

Council of Europe. (2008). *Spoken performances illustrating the 6 levels of the Common European Framework of References for Languages* [DVD]. Strasbourg: Language Policy Division.

Council of Europe. (2009*)*. Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment. A Manual. Strasbourg: Language Policy Division.

Council of Europe. (2011*)*. Manual for test development and examining: For use with the CEFR. Strasbourg: Language Policy Division.

Cizek, G. J., & Bunch, M. B. (2007). Standard setting: a guide to establishing and evaluating performance standards on tests. London: Sage Publications.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice (*Winter 2004).

Cohen, A. S., Kane, M.T., & Crooks, T. J. (1999). A generalized examinee-centered method for setting standards on achievement tests. *Applied Measurement in Education*, *12* (4), 343 – 366.

Haladyna, T. & Hess, R. (1999). An evaluation of conjunctive and compensatory standard-setting strategies for test decisions. *Educational Assessment, 6* (2), 129 – 153.

Hambleton, R. K., & Eignor, D. R. (1978, October). Competency test development, validation, and standard setting. Paper presented at the Minimum Competency Testing Conference of the American Education Research Association, Washington, DC. (Eric Documentation Reproduction Service No. ED 206 725).

Hambleton, R. K. & Pitoniak, M.J. (2006). Setting performance standards. In R. L. Brennan, *Educational Measurement - 4th edition* (pp.433 – 470). Westport, CT: American Council on Education and Praeger Publishers.

Hambleton, R.K., Pitoniak, M.J. & Copella, J.M. (2012). Essential steps in setting performance standards in educational tests and strategies for assessing the reliability of results. In G.J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations (pp. 47 – 76)*. New York: Routledge.

Hambleton, R.K. & Novick. M.R. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, 10 (3), 159 – 170.

Hanson, B.A., & Brennan, R.L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement*, *27* (4), 345 – 359.

Hellenic American University (2011). BCCE™ Test Administration Report: Revised BCCE™ Examination 2011. Available at http://www.hauniv.us/?i=hau-uni.en.downloads

Hyot, W. T. (2010). Interrater reliability. In G.R. Hancock & R.O. Mueller (Eds.), *The reviewer's guide to quantitative methods in social sciences* (pp. 141 - 154). New York: Routledge

McGraw K. O., & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1* (1), 30 – 46.

Irwin, P.M., Plake, B.S., & Impara, J.C. (2000). Validity of item performance estimates from an Angoff standard setting Study. Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000). Eric ED 443 875

Jaeger, R.M. (1991) Selection of judges for standard-setting. *Educational Measurement: Issues and Practice*, 10 (2), 3 – 14.

Kaftandjieva, F. (2004). *Standard Setting. Section B of the reference supplement to the preliminary pilot version of the manual for relating language examinations to the Common European Framework of Reference for languages: Learning, teaching, assessment*. Strasbourg: Language Policy Division, Council of Europe.

Kaftandjieva, F. (2010). *Methods for setting cut scores in criterion-referenced achievement tests: A comparative analysis of six recent methods with an application to tests of reading in EFL*. Arnhem: Cito.

Kaftandjieva, F., & Takala, S. (2002). Council of Europe scales of language proficiency: A validation study. In C.R. Alderson (Ed.), *Common European Framework of References for languages: Learning, teaching, assessment. Case studies*. (pp. 106 – 129). Strasburg: Council of Europe.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, *64*, 425 – 461.

Kane, M. (1998). Choosing between examinee–centered and test–centered standard–setting methods. *Educational Assessment*, *5* (3), 129 – 145.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classification based on test Scores. *Journal of Educational Measurement*, *32* (2), 179 – 197.

Maurer, T. J., & Alexander, R. A. (1992). Methods of improving employment test critical scores derived by judging test content: A review and critique. *Personnel Psychology*, *45* (4), 727 – 762.

Papageorgiou. S. (2010a). Linking international examinations to the CEFR: the Trinity College London experience. In W.Martyniuk (Ed.), *Aligning Tests with the CEFR: Reflections on using the council of Europe's draft Manual (pp*- 133 – 158). Cambridge: Cambridge University Press.

Papageorgiou, S. (2010b). *Setting cut scores on the Common European Framework of Reference for the Michigan English test: Technical report*. Cambridge Michigan Language Assessments. (http://www.cambridgemichigan.org/sites/default/files/resources/MET_StandardSetting.pdf)

Raymond, M. & Reid, J. (2001). Who made thee a judge? Selecting and training participants for standard setting, in Cizek, G. J. (Ed.) *Setting performance standards*. New Jersey: Erlbaum, 117 – 158.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86 (2)*, 420 – 428.

Stemler. S.E., & Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. In J.W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 29 – 49). California: Sage Publications, Inc.

Subkoviak, M. J. (1988). A practitioner's guide to computation of interpretation of reliability indices for mastery tests. *Journal of Educational Measurement*, *25* (1), 47 – 55.

Tanko, G. (2004). *Into Europe: The writing handbook*. Budapest: British Council.

Thompson, N.A., & Guyer, R. (2010). *User's Manual for Iteman 4 [Computer software manual]*. St. Paul MN: Assessment Systems Corporation.

Zieky, M.J., Perie, M. & Livingston, S.A. (2008). *Cutscores: A manual for setting standards of performance on educational and occupational tests*. Lexington: Educational Testing Service.

Appendices

Appendix 1: Confirmation E-mail Sent to Judges

Dear colleague:

You have been selected to serve on the **Basic Communication Certificate in English (BCCE™)** examination standard setting committee meeting in **Athens** on **July 26 – 29** of this year.

The task will be for you and selected educators/examiners to recommend a cut score to identify students who should receive a BCCE™ examination certificate. This will be a unique opportunity for you to participate in a decision that will significantly affect candidates and is likely to have important consequences for both local and international candidates.

I look forward to seeing you on **Tuesday, July 26 at 9:00 a.m**. (**Massalias 22 - 6th Floor Conference Room**), when you will register and pick up the materials you will need. Each session will involve working with CEFR descriptors and tasks and then evaluating a particular section of the BCCE™.

Please review the following: (1) the CEFR descriptors http://www.coe.int/T/DG4/Portfolio/documents/All%20scales%20CEFR.DOC; and (2) the CEFR tasks http://www.helsinki.fi/project/ceftrain/index.php.35.html

I look forward to seeing you on **Tuesday 26th July**.

Please confirm your attendance by replying to this email.

Should you need any further clarification, please do not hesitate to contact me.

Sincerely,

Charalambos Kollias
BCCE™ Examination Linking Project Coordinator

(adapted from Cizek & Bunch, 2007)

Appendix 2: Confidentiality Agreement Form

This agreement between _____ (**print committee member name**) and Hellenic American University provides for the review of test items for the Basic Communication Certificate in English (BCCE™) examination under the following terms and conditions:

1. The committee member will participate in training provided by Hellenic American University or designated subject-matter representatives for the purpose of setting standards for the BCCE™ examination.

2. The committee member will review test items in accordance with written specifications and verbal instructions provided by Office of Language Assessment and Test Development (OLATD) representatives.

3. The committee member will follow all test security procedures set forth in writing or verbally by Office of Language Assessment and Test Development (OLATD) representatives.

4. The committee member will turn over to Hellenic American University representatives all products of the standard setting meeting at the close of the session or as directed by Office of Language Assessment and Test Development (OLATD) representatives.

I understand that these test materials are restricted. I understand that all test questions and all other materials which are considered a part of the BCCE™ examination including, but not limited to, reading and listening passages, writing and speaking stimuli (prompts), grammar and vocabulary items, charts, graphs, and tables, are considered secure, and I will maintain complete confidentiality regarding all materials in these categories.

_____ 26/07/11

Committee member signature Date

_____ _____

Committee member e-mail address Committee member phone number

Charalambos Kollias 26/07/11

Date

Office of Language Assessment and Test Development Representative

Appendix 3: Examples of Familiarization Tasks

Example of Paper Familiarization Task

Match each descriptor with the appropriate CEFR Level (A1, A2, B1, B2, C1, C2)

Level	Descriptor
	1. Can generally identify the topic of discussion around him/her that is conducted slowly and clearly.
	2. Can follow the essentials of lectures, talks and reports and other forms of academic/professional presentation which are propositionally and linguistically complex.
	3. Can understand the main points of radio news bulletins and simpler recorded material about familiar subjects delivered relatively slowly and clearly.
	4. Can understand a wide range of recorded and broadcast audio material, including some non-standard usage, and identify finer points of detail including implicit attitudes and relationships between speakers.
	5. Can keep up with an animated conversation between native speakers.
	6. Can generally follow the main points of extended discussion around him/her, provided speech is clearly articulated in standard dialect.
	7. Can understand and extract the essential information from short recorded passages dealing with predictable everyday matters that are delivered slowly and clearly.
	8. Can easily follow complex interactions between third parties in group discussion and debate, even on abstract, complex unfamiliar topics
	9. Can follow specialised lectures and presentations employing a high degree of colloquialism, regional usage or unfamiliar terminology.
	10. Can understand instructions addressed carefully and slowly to him/her and follow short, simple directions.
	11. Can follow most lectures, discussions and debates with relative ease.
	12. Can understand simple technical information, such as operating instructions for everyday equipment.
	13. Can with some effort catch much of what is said around him/her, but may find it difficult to participate effectively in discussion with several native speakers who do not modify their language in any way.
	14. Can catch the main point in short, clear, simple messages and announcement.
	15. Can follow in outline straightforward short talks on familiar topics provided these are delivered in clearly articulated standard speech.
	16. Can follow a lecture or talk within his/her own field, provided the subject matter is familiar and the presentation straightforward and clearly structured.
	17. Can understand complex technical information, such as operating instructions, specifications for familiar products and services.
	18. Can understand announcements and messages on concrete and abstract topics spoken in standard dialect at normal speed.
	19. Can understand simple directions relating to how to get from X to Y, by foot or public transport.
	20. Can follow detailed directions.
	21. Can understand most radio documentaries and most other recorded or broadcast audio material delivered in standard dialect and can identify the speaker's mood, tone, etc.
	22. Can extract specific information from poor quality, audibly distorted public announcements e.g. in a station, sports stadium, etc.
	23. Can understand the information content of the majority of recorded or broadcast audio material on topics of personal interest delivered in clear standard speech.
	24. Can understand recordings in standard dialect likely to be encountered in social, professional or academic life and identify speaker viewpoints and attitudes as well as the information content.

Figure 3A Example of Online Familiarization Task

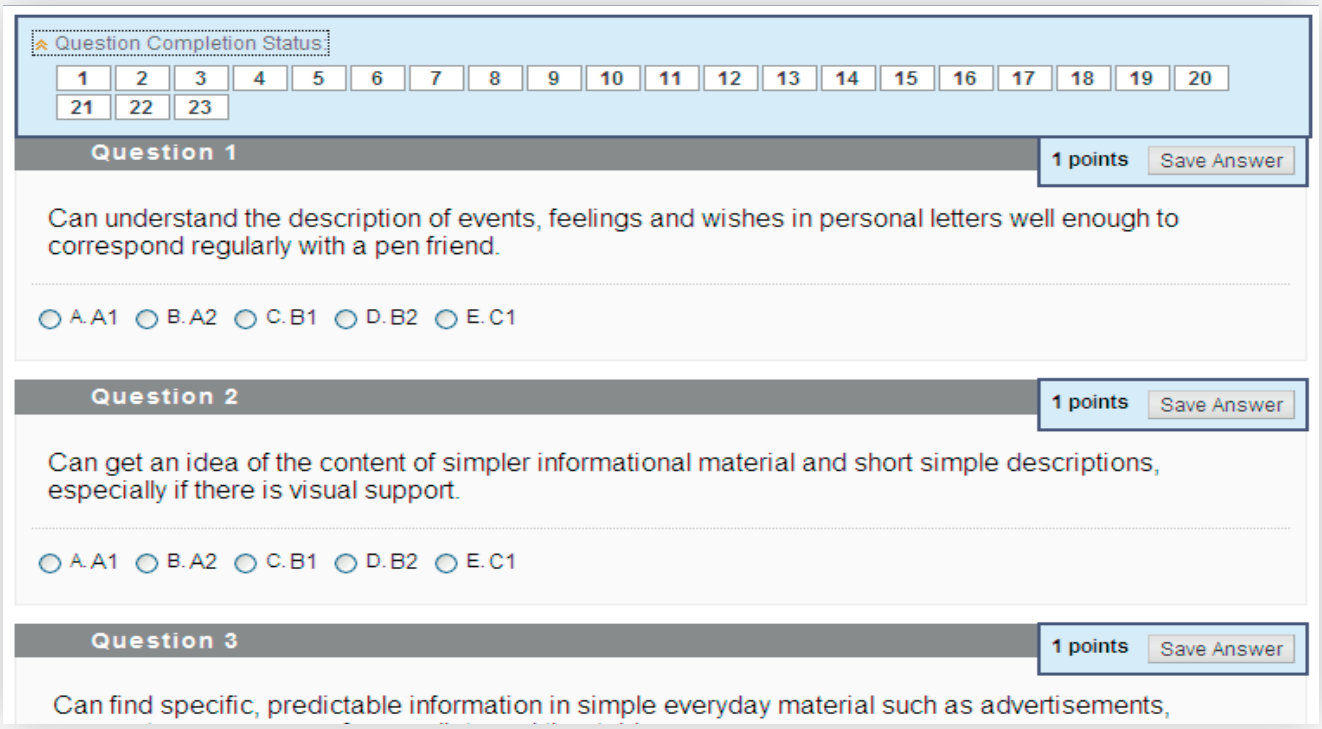


Figure 3B Example of Online Familiarization Task: Judge Feedback

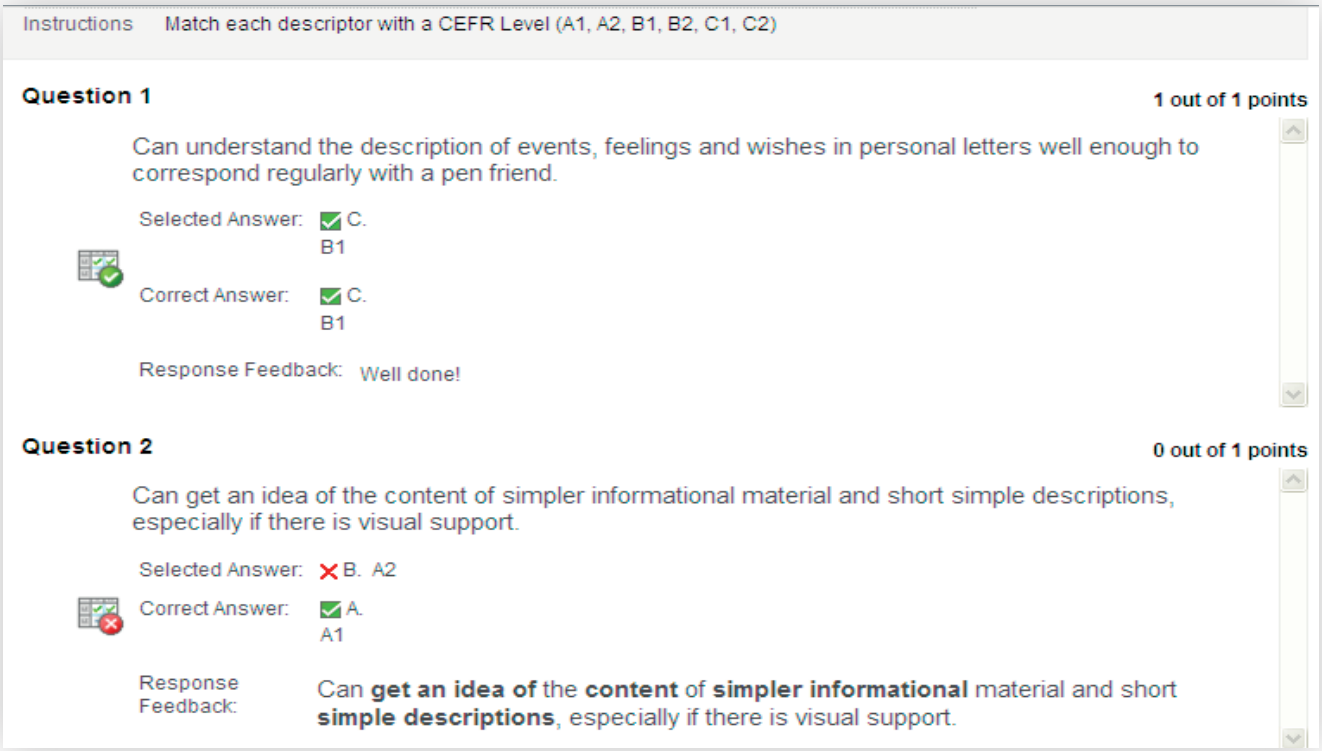
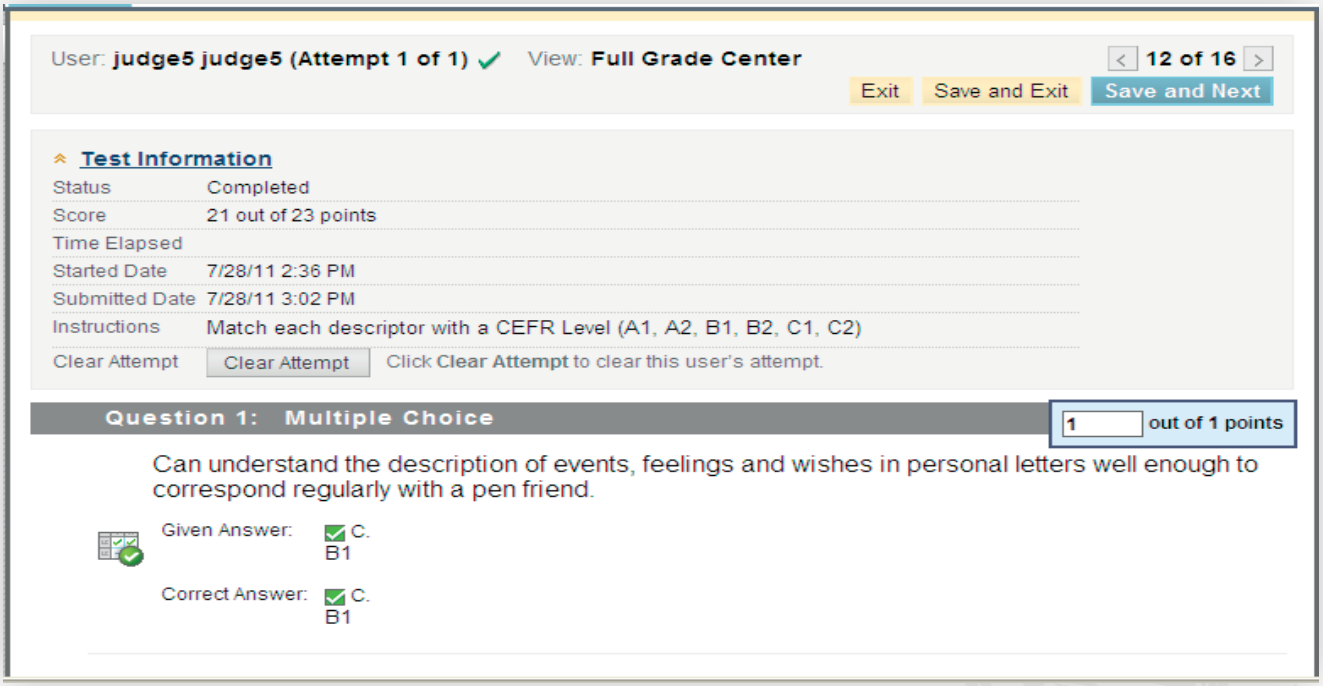


Figure 3C Example of Online Familiarization Task: Facilitator's Review



Appendix 4: Agenda for the BCCE™ Examination Standard Setting Workshop

Day 1	
9:00am - 9:15am	Registration
9:15am - 9:30am	Welcome and introductions
9:30am - 10:15am	Overview and orientation
10:15am - 10:30am	Break
10:30am - 10:45am	Online platform orientation
10:45am - 1:10pm	Speaking Familiarization
1:10pm - 2:10pm	Lunch
2:10pm - 3:30pm	Speaking Training (Illustration)
3:30pm - 3:45pm	Break
3:45pm - 4:35pm	Speaking Training (Controlled)
4:35pm - 4:45pm	Break
4:45pm - 5:15pm	Speaking Training (Freer)
5:15pm	Adjourn

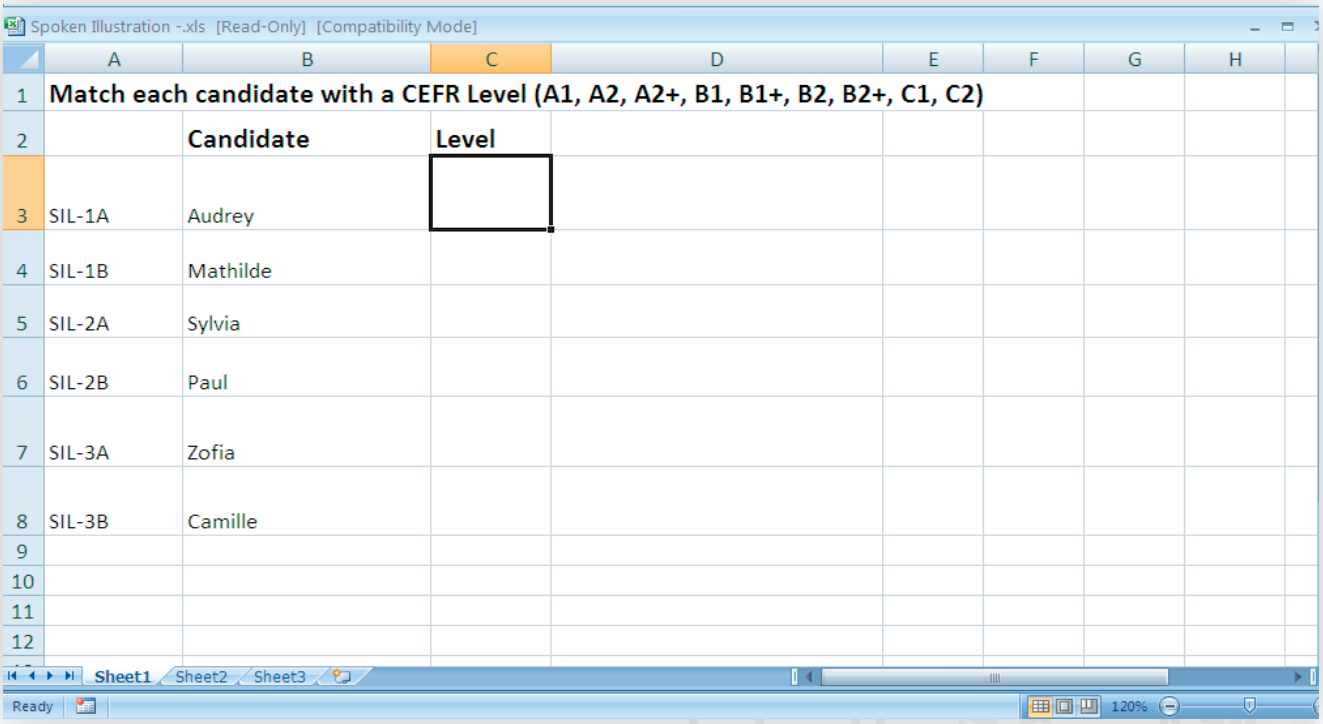
Day 2	
9:00am - 9:30am	Reception
9:30am – 10:00am	Speaking Training Summarization
10:00am – 11:05am	Speaking Benchmarking Locals L1 - L3
11:05 am - 11:20 am	Break
11:20am - 11:50am	Speaking Benchmarking Locals L1 - L3 cont.
11:50am- 1:15pm	Speaking Benchmarking Locals L4 - L6
1:15pm - 2:25pm	Lunch
2:25pm - 3:25pm	Speaking Benchmarking Locals L7 - L9
3:25pm - 3:40pm	Break
3:40pm - 5:00pm	Writing Familiarization
5:00pm	Adjourn

Day 3	
9:00am - 9:30am	Reception
9:30am - 11:25am	Writing Training
11:25am - 1:35pm	Writing - Benchmarking
1:35pm - 2:30pm	Lunch
2:30pm - 3:45pm	Reading, Grammar, & Vocabulary Familiarization
3:45pm - 4:15pm	Reading Training (Dialang)
4:15pm - 4:30pm	Break
4:30pm - 5:30pm	GVR section Timed
5:30pm	Adjourn

Day 4	
9:00am - 9:20am	Writing & Speaking Summarization of results & GVR Answers
9:20am - 9:45am	Modified Angoff method training
9:45am - 10:50am	GVR Round 1
10:50am - 12:00pm	GVR Round 1 Discussion
12:00pm - 12:25pm	GVR Reality Feedback
12:25 - 1:00pm	GVR Round 2
1:00pm - 1:45pm	Lunch
1:45pm - 2:00pm	GVR Summarization of results
2:00pm - 2:40pm	Listening Familiarization
2:40pm - 3:35pm	Listening Training
3:35pm - 3:50pm	Break
3:50pm - 4:25pm	Listening section Timed & Answers
4:30pm - 5:00pm	Listening Round 1
5:00pm - 5:30pm	Listening Round 1 Discussion
5:30pm - 5:45pm	Listening Round 2
5:45pm - 6:00pm	Listening Summarization of results
6:00pm - 6:15pm	Wrap Up

Appendix 5: Example of Speaking Section Training Rating Form

Figure 5 Example of Speaking Section Training Microsoft Excel® Rating Form



Appendix 6: Writing Section Round 2 Rating Form

Task1-01	
Task1-02	
Task1-03	
Task1-04	
Task1-05	
Task1-06	
Task1-07	
Task1-08	
Task1-09	
Task1-10	
Task1-11	
Task1-12	
Task1-13	
Task1-14	
Task1-15	
Task1-16	
Task1-17	
Task1-18	
Task1-19	
Task1-20	
Task1-21	
Task1-22	
Task1-23	
Task1-24	
Task1-25	

Task2-01	
Task2-02	
Task2-03	
Task2-04	
Task2-05	
Task2-06	
Task2-07	
Task2-08	
Task2-09	
Task2-10	
Task2-11	
Task2-12	
Task2-13	
Task2-14	
Task2-15	
Task2-16	
Task2-17	
Task2-18	
Task2-19	
Task2-20	
Task2-21	
Task2-22	
Task2-23	
Task2-24	
Task2-25	

Appendix 7: Information on P-values and R-pbis

The P value

The *P value* is the proportion of examinees that answered an item correctly (or in the keyed direction). It ranges from 0.0 to 1.0. A high value means that the item is easy, and a low value means that the item is difficult. The *minimum P value* bound represents what you consider the cut point for an item being **too difficult**. For a relatively easy test, you might specify 0.50 as a minimum, which means that 50% of the examinees have answered the item correctly. For a test where we expect examinees to do poorly, the minimum might be lowered to 0.4 or even 0.3. The minimum should take into account the possibility of guessing; if the item is multiple-choice with four options, there is a 25% chance of randomly guessing the answer, so the minimum should probably not be 0.20. The *maximum P value* represents the cut point for what you consider to be an item that is **too easy**. The primary consideration here is that if an item is so easy that nearly everyone gets it correct, it is not providing much information about the examinees. In fact, items with a *P* of 0.95 or higher typically have very poor point-biserial correlations.

The item point-biserial (r-pbis) correlation

The point-biserial correlation (*r-pbis*) is a measure of the discrimination, or differentiating strength, of the item. It ranges from -1.0 to 1.0. A good item is able to differentiate between examinees of high and low ability, and will have a higher point-biserial, but rarely above 0.50. A negative point-biserial is indicative of a very poor item, because then the high-ability examinees are answering incorrectly, while the low examinees are answering it correctly. A point biserial of 0.0 provides no differentiation between low-scoring and high-scoring examinees, essentially random –noise. The *minimum r-pbis* bound represents the lowest discrimination you are willing to accept. This is typically a small positive number, like 0.10 or 0.20. If your sample size is small, it could possibly be reduced. The *maximum r-pbis* bound is almost always 1.0, because it is typically desired that the *r-pbis* be as high as possible.

Source: Thompson & Guyer (2010)

Appendix 8: Listening Section Data Entry Microsoft Excel® Worksheets

Figure 8A Listening Section Microsoft Excel® Round 1 Worksheet

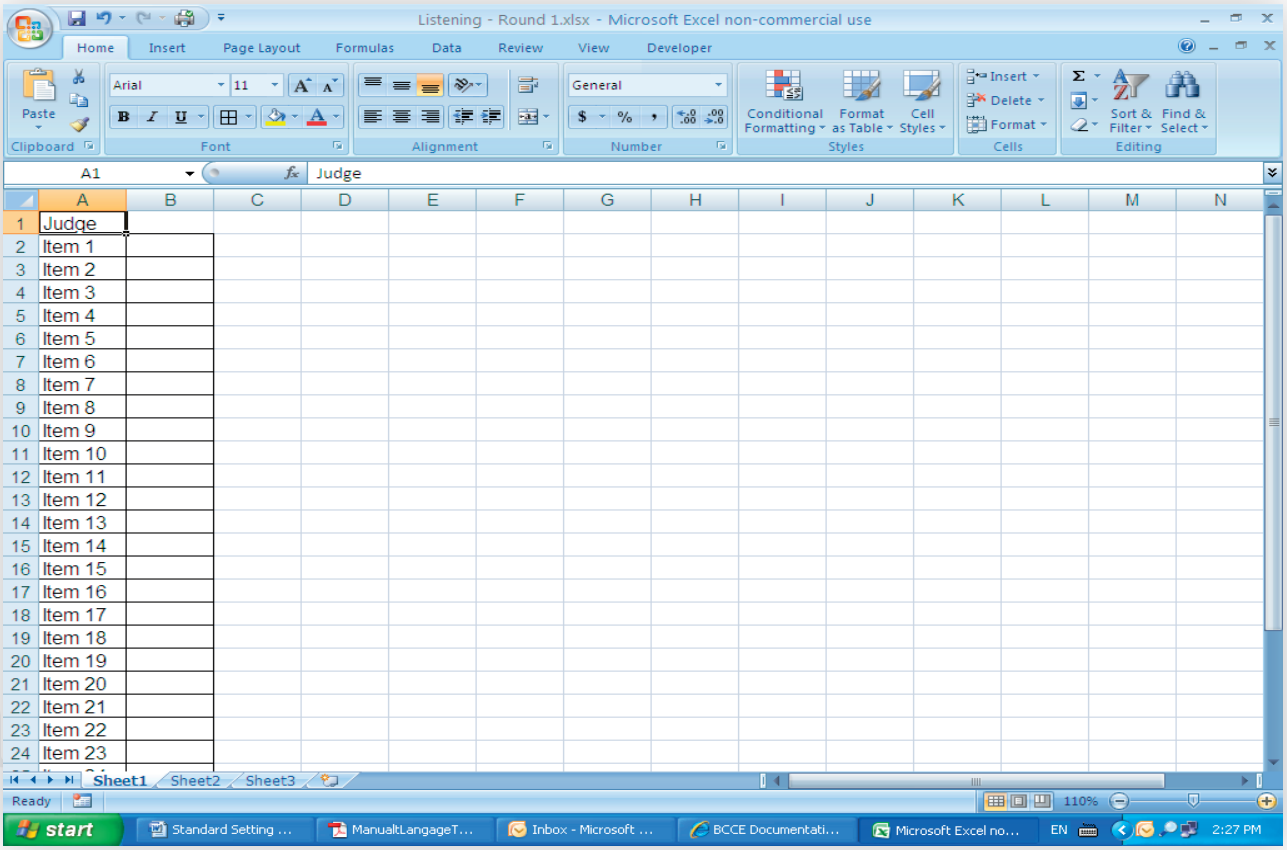
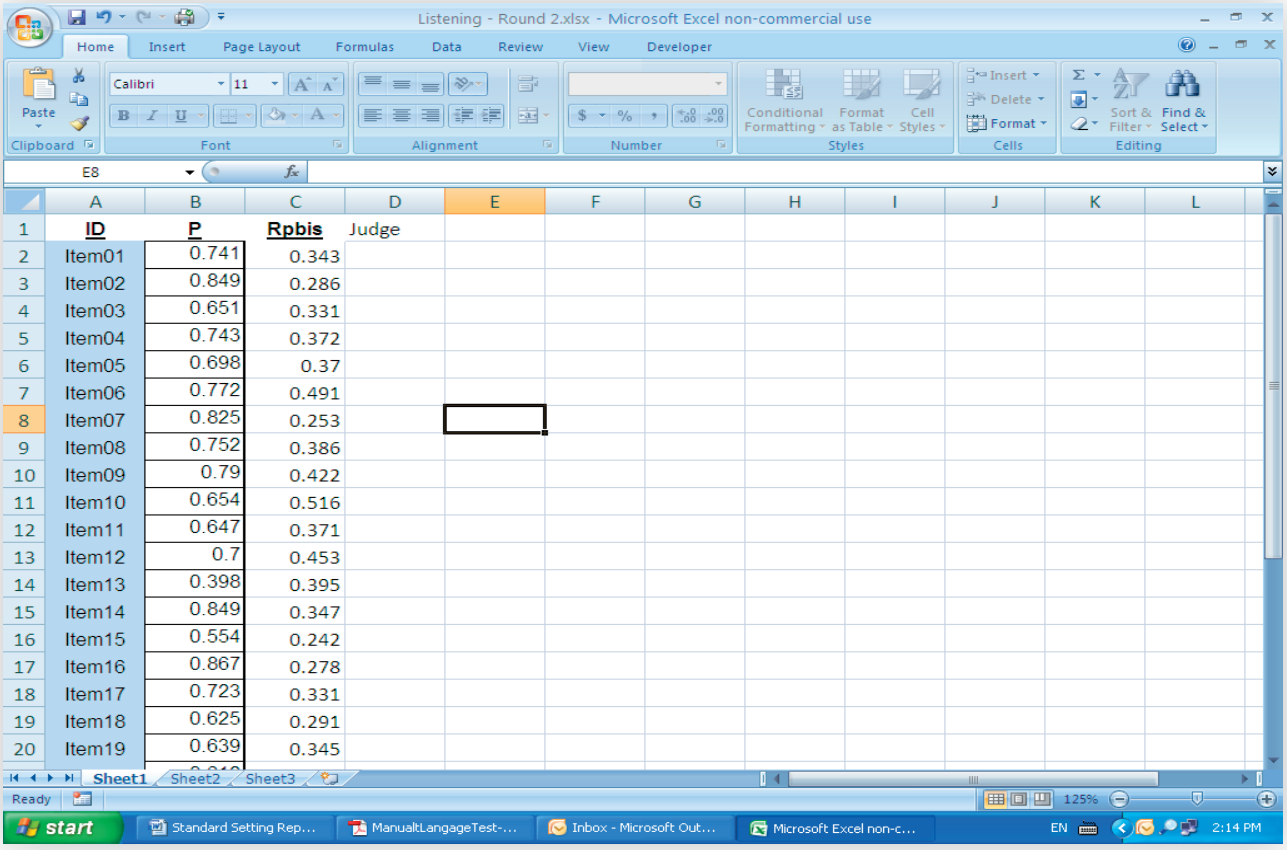


Figure 8B Listening Section Microsoft Excel® Round 2 Worksheet



Office for Language Assessment

505 Amherst Street, Nashua,

New Hampshire, NH 03063, USA

t: +1 603-577-8700

e-mail: info@hauniv.edu