2010–2013
Technical Review

CONTACT INFORMATION

All correspondence and mailings should be addressed to:

CaMLA
Argus 1 Building
535 West William St., Suite 310
Ann Arbor, Michigan
48103-4978 USA

T: +1 866.696.3522
T: +1 734.615.9629
F: +1 734.763.0369

info@cambridgemichigan.org
CambridgeMichigan.org

10/2014

Table of Contents

List of Tables

1. Introduction

The Examination for the Certificate of Competency in English (ECCE) is a test of general language proficiency for learners of English. From 2010 to 2013, the exam was administered nine times, a minimum of twice per year, at test centers around the world.

This report provides test users with technical information about the ECCE. Section 2 provides general information about the test and a proposed interpretation of ECCE test scores. In Section 3, the report explains how the exam is scored and equated, and the procedures for reporting scores. It also gives guidelines for score use in decision making. Section 4 describes the changes in the ECCE from 2010 to 2013. Section 5 discusses the ECCE test taking population, looking particularly at the yearly distributions of test takers by gender and age. Section 6 looks at trends in the ECCE test results by age and gender. It also examines trends in reliability estimates, standard error of measurement, and subtest correlation for each year. The final section of the report reviews the validity evidence currently available to support CaMLA's proposed interpretation of the ECCE results.

2. Description of the ECCE

2.1. General Description

The ECCE is a standardized high-intermediate level English as a foreign language (EFL) examination. It is a test of general language proficiency in a variety of contexts. The four component skills of listening, reading, writing, and speaking are evaluated through a combination of tasks.

The ECCE is aimed at the B2 level of the Common European Framework of Reference (CEFR) and is valid for the lifetime of the recipient. The ECCE certificate is recognized in several countries as official documentary evidence of high-intermediate competence in English for personal, public, educational, and occupational purposes.

CAMLA is committed to the excellence of its tests, which are developed in accordance with the highest standards in educational measurement. All parts of the examination are written following specified guidelines, and items are pretested to ensure that they function properly. CaMLA works closely with test centers to ensure that its tests are administered in a way that is fair and accessible to test takers and that the ECCE is open to all people who wish to take the exam, regardless of the school they attend or their participation in formal language study.

2.2. Proposed Interpretation of the Scores

The ECCE is aimed at the B2 proficiency level on the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (Council of Europe, 2001) (CEFR). Language users at this level:

> Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialization. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
>
> (Council of Europe, 2001: 24)

Therefore, ECCE certificate holders can be expected to understand conversations and discussions in all areas of their social, professional, and academic life. If a lecture is particularly complex (in content and language) they will be able to grasp the main ideas. They can speak on a variety of topics, elaborating on their ideas and providing examples. They can also read a wide range of texts, varying their reading speed and focus to their reading purpose. Their vocabulary is flexible and they can infer the meaning of words from context.

2.3. Test Structure

The ECCE tests all four skill areas: listening, reading, writing, and speaking. Table 2.3 describes the format and content of the ECCE. Sample items for each section of the test are available on the CaMLA website.

Table 2.3 Format and Content of the ECCE

Section	Time	Description	Number of Items
Speaking	15 minutes	Test takers participate in a structured, multistage task with one examiner.	4 stages
Listening	30 minutes	**Part 1** (multiple choice) A short recorded conversation is followed by a question. Answer choices are shown as pictures.	30
		Part 2 (multiple choice) Short talks delivered by single speakers on different topics, followed by 4 to 6 questions each.	20
Grammar Vocabulary Reading	90 minutes	**Grammar** (multiple choice) An incomplete sentence is followed by a choice of words or phrases to complete it. Only one choice is grammatically correct.	35
		Vocabulary (multiple choice) An incomplete sentence is followed by a choice of words or phrases to complete it. Only one word has the correct meaning in that context.	35
		Reading (multiple choice) Part 1: A short reading passage is followed by comprehension questions. Part 2: Two sets of four short texts related to each other by topic are followed by 10 questions each.	30
Writing	30 minutes	The test taker reads a short excerpt from a newspaper article and then writes a letter or essay giving an opinion about a situation or issue.	1 task

3. Scoring and Reporting of ECCE Results

3.1. Explanation of Scoring for Each Section

The listening and grammar, vocabulary and reading (GVR) sections of the ECCE are scored by computer at CaMLA. Each correct answer carries equal weight within each section and there are no points deducted for wrong answers. A scaled score, ranging from 0 – 1000, is calculated using an advanced mathematical model based on Item Response Theory (IRT). This method ensures that the ability required to pass a section, or to receive a high score, remains the same from year to year.

The speaking section is conducted and assessed by a certified speaking examiner. The writing section is assessed by specialized raters trained and certified according to our standards. All essays are scored by at least two raters. The speaking and writing sections are graded according to scales established by CaMLA (see the CaMLA website for the ECCE speaking and writing rating scales). During the period covered by this report the speaking and writing sections were scored holistically; the test takers received a band score (A – E) for each section.

If a test taker's scores (the rating for speaking and writing; the scaled score for listening and GVR) meet the cutoff level in a section, they are given a pass for that section of the exam. During the period covered by this report test takers who passed three sections with a Low Pass (or higher) and received no less than a Borderline Fail in one section were awarded an ECCE certificate.

3.2. Equating Procedures

In order to ensure that ECCE scores obtained from different test forms are comparable and that fair decisions can be made regarding test results, the process of common item equating is used. Link items on each exam serve as the common items that are used to equate the different exam forms using item difficulty. Item difficulties from previous administrations are stored in a database. When established items are used as link items, their difficulty in the previous

administration is correlated with their difficulty in the current administration. This enables CaMLA to calculate equated scale and location parameters. These parameters allow different forms of the ECCE to be equated. The scale and location parameters are computed separately for the listening and GVR sections and are implemented in a scoring run in BILOG-MG.

3.3. Procedures for Reporting Results

All test takers receive an Examination Report that shows their overall performance as well as the levels for each test section. Test takers are given these results so that they will know the areas in which they have done well and those in which they need to improve. The Examination Report provides the following information:

- The result for the ECCE (Pass/Fail)
- Section results with a brief description of the test taker's performance.

ECCE section scores are reported in five bands (see Table 3.3). The score report also provides a numeric score for the listening and GVR sections. The numeric score provides test takers with more precise information on their performance. For example, a test taker who receives a band score of Pass (P) in the listening section of the ECCE can see if his or her score is at the top of this band, closer to High Pass (HP), or if it is closer to a Low Pass (LP).

Table 3.3 ECCE Performance Range

Score Band	Writing & Speaking	Listening & GVR
High Pass (HP)	A	840–1000
Pass (P)	B	750–835
Low Pass (LP)	C	650–745
Borderline Fail (BF)	D	610–645
Fail (F)	E	0–605

3.4. Interpretation of Scores for Each Section

As stated in the description of the exam (Section 2.2), the ECCE is aimed at the B2 (Vantage) level of the CEFR. Test takers who achieve a minimum band score of "C" in each section can be expected to have the following skills and abilities:

Speaking They are able to make clear, detailed presentations on subjects related to their field, providing supporting detail and highlighting significant points. In a discussion, they can explain the advantages and disadvantages of options and can support their views. They can interact with a degree of spontaneity and are generally able to turn-take such that it is possible to have a smooth flowing conversation.

Writing On subjects related to their interests, they are able to provide information and give reasons in support of a particular point of view. They are also able to express their feelings about events and experiences.

Listening If materials are presented in a standard dialect, they are able to understand most television news and current affairs programs as well as films. They can also understand extended conversations and lectures, even when the discussion is complex, as long as the topic is familiar.

Reading They are able to read current affairs articles and reports and can infer the writers' attitudes or viewpoints. They can also understand contemporary fiction. They can follow complex instructions as long as the topic is familiar.

Use of English They have a sufficient range of vocabulary and grammatical structures to convey their meaning on topics related to their field or general interest. They are able to use paraphrase even when they do not know the precise word for a concept. They can also monitor their language and correct their errors, including resolving misunderstandings.

3.5. Guidelines for Decision-Making

When interpreting an ECCE score report, it is important to remember that the ECCE estimates the test taker's true proficiency by approximating the kinds of tasks that they may encounter in real life. Also, temporary factors unrelated to a test taker's English

proficiency, such as fatigue, anxiety, or illness, may affect exam results.

When using test scores for decision-making, check the date the test was taken. While the certificate is valid for a holder's lifetime, language ability changes over time. This ability can improve with active use and further study of the language, or it may diminish if the holder does not continue to study or use English on a regular basis. It is also important to remember that test performance is only one aspect to be considered. Communicative language ability consists of both knowledge of language and knowledge of the world. Therefore, one would need to consider how factors other than language affect how well someone can communicate. For example, in the general context of using English in business, the ability to function effectively involves not only knowledge of English, but also other knowledge and skills such as intellectual knowledge and other business skills.

4. Changes to the ECCE from 2010–2013

During the period covered by this report CaMLA introduced three major changes to the ECCE. One was a scoring change and two were changes to the design of the test.

4.1. Changes to Test Design

During 2011 and 2012, CaMLA completed a review of the ECCE listening and GVR sections in order to ensure the continued excellence of the exam. This process involved a review of the current language assessment literature, a re-examination of the CEFR listening and reading scales, analyses of test taker performance, reviews of current item types, and investigations of alternative listening and reading tasks. Based on this comprehensive review, the revised listening and GVR sections were introduced in May 2013.

Revised Listening Section

Prior to May 2013, part 2 of the listening section consisted of 20 questions about a radio interview related to a single event. The interview was played in segments, and the questions were delivered orally, after each segment. The listening section revision resulted in a new listening task, short monologues, that would replace the previous task. Test takers hear four short talks, each about a minute and a half long that are each followed by four to six questions. The questions and answer options are printed in the text booklet.

Revised GVR Section

The GVR section revision resulted in several changes to the reading tasks. Beginning with the May 2013 administration, the advertisement and longer related passage task types were retired and a new reading task type was added. This task consists of 10 questions about four short reading texts (about 550 words) related to each other by topic. Additionally, the short reading passage task type previously on the ECCE was expanded to include two short passages, rather than one. As a result of these changes the time allowed for the GVR section was also increased from 80 minutes to 90 minutes.

4.2. Scoring Changes

Score Reporting

Beginning with the May 2011 administration, the ECCE score report has included not only the band score for each section (i.e. high pass, borderline fail, etc.) but also a scaled numeric score (0 – 1000) for the listening and GVR sections. This information was added to provide examinees with more precise information about their performance.

5. ECCE Test-Taking Population

This section presents an overview of the test takers who took the ECCE during the period covered by this report, providing information about the test takers' first languages, as well as the distributions for test taker gender and age.

5.1. First Language

Every ECCE test taker completes a registration form that asks for their first language. Cases where information is not given, or is not correctly given, are treated as missing data. Table 5.1 lists the first language backgrounds for the test takers who took the test during the period 2010–2013. It shows that the ECCE was taken by test takers from 51 different first language backgrounds.

Table 5.1 ECCE Test Taker First Languages

Afrikaans	Estonian	Lithuanian
Albanian	Fanti	Marathi
Amharic	Farsi/Persian	Norwegian
Arabic	French	Polish
Armenian	Georgian	Portuguese
Bambara/Malinke	German	Romanian
Bangali	Greek	Russian
Bulgarian	Gujarati	Slovak
Cambodian	Hausa	Spanish
Catalan	Hebrew	Swahili
Creole	Hindi	Swedish
Croatian	Hmong	Talalog/Filipino
Czech	Ibo (Igbo)	Turkish
Dari	Italian	Ukrainian
Dutch	Japanese	Urdu
Efik	Kannada	Vietnamese
English	Korean	Yoruba

5.2. Gender Distribution

The ECCE registration form also asks for the test taker's gender. Cases where information is not given, or is not correctly given are treated as missing data. Table 5.2 shows the distribution of test takers by gender from 2010 to 2013. The distribution is very similar from year to year; the data shows that just over half of the ECCE candidature tends to be female.

Table 5.2 Distribution (in %) of ECCE Test Takers by Gender

Year	Male (%)	Female (%)	Missing Data (%)
2010	45.38	54.53	0.08
2011	45.25	54.69	0.05
2012	45.21	54.68	0.11
2013	46.01	53.87	0.11

5.3 Age Distribution

The ECCE registration form also asks for the test taker's date of birth. As with first language and gender, cases where information on date of birth is not given, or is not correctly given are treated as missing data. Table 5.3 presents the distribution of test takers by age for each year. The distributions are very similar from year to year. The table shows that an overwhelming number of test takers were between 13 and 16 years old. This suggests that the majority of the ECCE candidates are adolescents who register for the exam while still in secondary school.

Table 5.3 Distribution (in %) of ECCE Test Takers by Age

Age	2010	2011	2012	2013
<12	2.16	2.36	2.63	2.81
13–16	79.91	80.13	80.12	80.79
17–19	6.32	6.28	6.09	5.74
20–22	4.49	4.29	4.03	3.61
23–25	2.98	2.91	2.77	2.75
26–29	1.86	1.75	1.88	1.65
30–39	1.47	1.41	1.58	1.56
>40	0.75	0.80	0.86	1.00
Missing Data	0.06	0.06	0.05	0.09

6. ECCE Results and Test Statistics

6.1. Trends in Overall Scores and Pass Rates for Individual Sections

Table 6.1.1 shows the distribution of ECCE overall pass scores from 2010 to 2013. The table reveals an upward trend in the overall pass rate. Since 2010, the pass rate has gradually increased by approximately 3 percentage points each year, which has resulted in an increase of just over 8 percentage points to the ECCE's overall pass rate during this time period.

Table 6.1.1 Overall Pass Rate of the ECCE

Year	Pass	Fail
2010	64.84	35.16
2011	67.06	32.94
2012	70.37	29.63
2013	72.90	27.10

Because the ECCE is carefully monitored to ensure that each test form is aimed at the B2 level of the Common European Framework of Reference (CEFR), this rise in the pass rate suggests that test takers are increasingly better prepared for the examination. This hypothesis

is further examined through analysis of the pass rate trends for each section of the ECCE.

Table 6.1.2 presents the pass rates for each section of the ECCE over the four year time period. This table shows that the pass rate for each section has gradually increased in the period 2010–2013. The increase is particularly notable for the listening and GVR sections (a rise of approximately 7 and 8 percentage points respectively). This uniform increase in pass rate across the sections supports the hypothesis that the test takers are increasingly better prepared for the exam. These trends are analyzed in more detail through the examination of the score distributions for each section of the ECCE.

Table 6.1.2 Pass Rates for Each Section of the ECCE

Year	Listening	GVR	Writing	Speaking
2010	68.60	63.37	82.01	86.25
2011	70.72	65.32	83.43	87.75
2012	71.43	70.79	85.84	88.60
2013	75.61	71.53	86.45	89.63

Listening

Table 6.1.3 shows the distribution (in %) of scores from 2010 to 2013 for the listening section of the ECCE. Even though the distribution of test takers across the performance bands (High Pass – Fail) has varied, the listening section's pass rate has been increasing steadily in recent years (see Table 6.1.2). There could be a number of reasons for this trend. One is that the candidature is becoming more able. Performance on common items, used as part of the ECCE equating procedure, suggests that this is a plausible explanation. Bearing in mind that the revised listening section was introduced in 2013, it is interesting to see that this coincided with a dramatic increase (of over 10 percentage points) in the number of High Pass results awarded. It is possible that the revised listening section gives the more able test takers more opportunities to express the full range of their listening abilities.

GVR

Table 6.1.4 shows the distribution (in %) of scores from 2010 to 2013 for the GVR section of the ECCE. As with the listening section, there is an upwards trend in the % of passes (see Table 6.1.2). Like the listening section, one explanation could be that the candidature is becoming more able. Performance on the GVR common items suggests that this is a plausible conclusion.

Table 6.1.3 Distribution (in %) of Scores on the ECCE Listening Section

Year	High Pass	Pass	Low Pass	Borderline Fail	Fail
2010	4.71	40.46	23.44	14.26	17.15
2011	5.64	23.70	41.38	12.44	16.84
2012	5.02	24.57	41.84	12.97	15.60
2013	16.59	27.72	31.30	8.75	15.64

Table 6.1.4 Distribution (in %) of Scores on the ECCE GVR Section

Year	High Pass	Pass	Low Pass	Borderline Fail	Fail
2010	9.86	34.69	18.82	13.09	23.55
2011	5.90	20.71	38.72	13.41	21.27
2012	10.72	23.68	36.39	11.37	17.84
2013	7.69	24.87	38.98	11.91	16.56

Writing

Table 6.1.5 shows the distribution (in %) of scores from 2010 to 2013 for the writing section of the ECCE. Analysis of this distribution reveals that it is very consistent from year to year: there is a slight increase in the pass rate year on year. The pass rate for the writing section is approximately 10 percentage points higher than that of the listening and GVR sections (see Table 6.1.2). Interestingly, the distribution of test takers across the performance bands is also markedly different; the majority of test takers receive low pass scores, and very few receive high pass or fail scores (approximately 1% each). This suggests that raters avoid awarding the extreme categories of the rating scale. Because the overall score of the ECCE allows for a single borderline fail, underuse of the fail category in the writing section might be contributing to the upward trend in the ECCE overall pass rate. This pattern can partially be explained by the fact that test takers will register for the ECCE when they consider themselves minimally able to pass the exam (there being no particular advantage to them waiting until they are able to score well on the exam). Nevertheless, the sheer numbers of test takers who are awarded a Low Pass suggests a strong tendency for convergence towards the mean (cf., Leckie & Baird, 2011). This possibility needs further investigation and will need to be addressed in rater training.

Speaking

Table 6.1.6 shows the distribution (in %) of scores from 2010 to 2013 for the speaking section of the ECCE. Like the writing section, the speaking section score distribution changes very little from year to year. There is a slight increase in the pass rate year on year and the pass rate is 14 – 18 percentage points higher than for the listening and GVR sections (see Table 6.1.2). From Table 6.1.6 we can also see that the fail category again appears to be underused; a very small percentage of test takers obtain this score. As in the case of the writing section, this is a source of concern: it indicates that scores for the speaking section of the ECCE are out of line with the listening and GVR sections. Also, as we have suggested for the writing section, since the overall score for the ECCE allows for a single borderline fail, underuse of the fail category in the speaking test might be a contributing factor to the upwards trend in the pass rate for the ECCE overall.

6.2. Distribution of Results by Age and Gender

Age

Table 6.2.1 presents the percentage of test takers in each age band who received an overall pass on the ECCE each year. It is important to note that the percentage of the test-taking population younger than 13 and above 22 years of age is very small (see Table 5.3, above), so the results for those groups should be

Table 6.1.5 Distribution (in %) of Scores on the ECCE Writing Section

Year	High Pass	Pass	Low Pass	Borderline Fail	Fail
2010	0.93	14.03	67.05	16.62	1.37
2011	1.20	14.74	67.49	15.40	1.17
2012	1.05	12.67	72.12	13.09	1.07
2013	1.22	17.25	67.98	12.22	1.34

Table 6.1.6 Distribution (in %) of Scores on the ECCE Speaking Section

Year	High Pass	Pass	Low Pass	Borderline Fail	Fail
2010	14.13	30.73	41.38	11.37	2.38
2011	13.61	29.82	44.32	10.48	1.77
2012	12.11	29.75	46.73	10.14	1.26
2013	12.97	30.50	46.17	9.29	1.07

interpreted with caution. The data suggests that there is a small age effect on the exam's pass rate. This effect is most marked when comparing the largest age groups (13–16 and 17–19). The much lower pass rates of the 17–19 year olds suggests that they are less proficient than the 13–16 year olds or perhaps not as well-prepared for the exam.

Table 6.2.1 Percentage of Test Takers for Each Age Group Who Received an Overall Pass

Age	2010	2011	2012	2013
<12	80.88	82.96	83.98	86.84
13–16	65.09	67.52	70.64	73.42
17–19	58.32	61.57	64.59	66.24
20–22	59.64	60.82	67.57	66.44
23–25	66.03	63.90	65.10	65.91
26–29	66.18	66.23	71.47	68.26
30–39	70.05	70.04	73.61	75.73
>40	60.13	60.75	66.06	70.75

In order to establish whether these differences were meaningful, we ran cross-tabulations and chi-square tests for each year. Table 6.2.2 summarizes the Pearson Chi-Square value (χ^2) for each section, as well as the degrees of freedom (df), the level of significance (p), and a measure of effect size, Cramer's V. Cramer's V provides a measure of the strength (meaningfulness) of the association between two variables, taking account of sample size and degrees of freedom (Field, 2005: 692). It produces a value between 0 and 1, where higher values indicate stronger association.

Table 6.2.2 Chi-Square Test Results for Age and ECCE Pass Rate

Year	χ^2	df	p	Cramer's V
2010	276.13	7	<0.001	0.067
2011	268.09	7	<0.001	0.069
2012	206.63	7	<0.001	0.063
2013	295.44	7	<0.001	0.078

The table shows that there was a significant association between a test taker's age and whether they received an overall pass for the ECCE. However, the Cramer's V measure indicates that the effect size is small; that is, the association between age and overall pass rate may not be sufficiently large to be meaningful. On the basis of these analyses, it is not possible to claim that any particular age group is more (or less) likely to pass the ECCE than any other group.

Gender

Table 6.2.3 presents the percentage of male and female test takers who received an overall pass on the ECCE each year. The data suggests that male test takers tend to perform better on the ECCE than female test takers. As in the case of the age groups, we ran cross-tabulations and chi-squared tests for each year in order to determine whether the association was meaningful.

Table 6.2.3 Percentage of Test Takers for Each Gender Who Received an Overall Pass

Gender	2010	2011	2012	2013
Male	67.45	69.62	72.67	74.73
Female	62.66	64.93	68.44	71.34

Table 6.2.4 summarizes the Pearson Chi-Square value (χ^2) for each section, as well as the degrees of freedom (df), the level of significance (p), and a measure of effect size, Cramer's V. It shows that, for each year, there was a significant association between the gender of the test taker and whether they received an overall pass for the ECCE. However, the Cramer's V measure indicates that the effect size is small; that is, the association between gender and overall pass rate may not be sufficiently large to be meaningful. On the basis of these analyses, it is not possible to claim that male test takers are more likely to pass the ECCE than female test takers.

Table 6.2.4 Chi-Square Test Results for Gender and ECCE Pass Rate

Year	χ^2	df	p	Cramer's V
2010	154.52	1	<0.001	0.050
2011	139.79	1	<0.001	0.050
2012	108.99	1	<0.001	0.046
2013	69.93	1	<0.001	0.038

6.3. Trends in Reliability Estimates and Rater Agreement Statistics

Test scores are a numerical measure of a test taker's ability. *Reliability* refers to the consistency of that measurement. In theory, a test taker's test score should

be the same each time the test is taken or across different forms of the same test. In practice, even when the test conditions are carefully controlled, an individual's performance on a set of test items will vary from one test administration to another due to variation in the items across different forms of the same test or due to variability in individual performance. Among the reasons for this are temporary factors unrelated to a test taker's proficiency, such as fatigue, anxiety, or illness. As a result, test scores always contain a small amount of measurement error. The aim is to keep this error to a minimum. For high-stakes exams such as the ECCE, which aim at a specific proficiency level, a reliability figure of 0.80 or above is considered acceptable.

Table 6.3.1 Reliability Estimates for the ECCE Listening and GVR Sections

Year	Listening	GVR
May 2010	0.89	0.93
December 2010	0.87	0.94
May 2011	0.88	0.94
December 2011	0.87	0.94
May 2012	0.89	0.94
December 2012	0.88	0.94
May 2013	0.89	0.95
September 2013	0.87	0.91
December 2013	0.89	0.95

Reliability estimates are obtained for the listening and GVR sections on each administration of the ECCE. They are calculated with the program, BILOG, using the Bayes MAP (maximum *a posteriori*) method. Table 6.3.1 presents the reliability estimates for each ECCE administration in the period 2010–2013. It shows that the reliability estimates for the listening section is typically lower than that of the GCVR section. This is probably because of the relative length of the sections; the listening section comprises 50 items whereas the GCVR section comprises 100 items. Nevertheless, both sections are consistently well above the acceptable value of 0.80. The estimates are also similar for each administration, which suggests an excellent consistency of measurement in the ECCE.

In the case of performance tests such as the writing and speaking sections of the ECCE, the reliability of the score awarded can be affected by the consistency of the rating process. For this reason, it is also important to monitor these sections. The examiners for the speaking test are native or highly proficient nonnative speakers of English who are trained and certified according to standards set by CaMLA. The examiner who conducts the speaking test assesses and rates the test taker's performance using the ECCE Speaking Rating Scale. Because the ECCE speaking test is administered by only one examiner, it is not possible to obtain rater agreement figures. Instead, performance of speaking test examiners is monitored locally by senior experienced rater training examiners and recordings of speaking tests are sent to CaMLA for review.

The raters for the writing section are native speakers of English, all trained and certified according to standards set by CaMLA. Each writing performance is rated separately by two accredited raters. If these raters do not reach exact agreement on the score to be awarded, the writing performance is evaluated separately by a third rater. It is important to note that *the final score awarded for each ECCE writing section response is the result of exact agreement by a minimum of two raters who have each independently evaluated the writing performance*. This means that no single rater can determine the final outcome for a script.

CaMLA monitors rater agreement for training purposes. It monitors exact agreement between the first and second rater. It also monitors pass/fail agreement; that is, the extent to which raters agree on whether a performance should be awarded a passing grade or a failing grade. Table 6.3.2 presents these rater agreement figures for each administration of the writing section.

Table 6.3.2 Rater Agreement Figures for the Writing Section

Year	Rater 1 / Rater 2 Agreement (%)	Pass/Fail Agreement (%)
May 2010	75.90	89.60
December 2010	77.80	88.90
May 2011	69.90	88.00
December 2011	74.06	86.94
May 2012	76.80	89.90
December 2012	72.80	88.30
May 2013	72.45	90.38
September 2013	81.71	96.57
December 2013	78.01	91.10

The table shows that both the exact rater agreement and the pass/fail agreement are at good levels. While

the exact rater agreement varies a little between administrations, it is generally around 75%. The pass/fail agreement is more stable, and is generally around 90%. CaMLA is constantly working with the raters for the writing section to maintain and improve these agreement figures.

6.4. Trends in Standard Error

Apart from monitoring the reliability estimates, the estimated variability in test taker performance can also be monitored through the standard error of measurement (SEM) estimates. As mentioned in Section 6.3, test scores always contain a small amount of measurement error. The aim is to keep this error to a minimum.

Table 6.4 SEM Estimates for the ECCE Listening and GVR Sections

Year	Listening	GVR
May 2010	34.00	27.00
December 2010	36.00	25.00
May 2011	35.00	24.00
December 2011	34.00	23.00
May 2012	32.85	24.47
December 2012	34.53	23.62
May 2013	33.08	22.76
September 2013	36.37	30.30
December 2013	33.30	23.30

SEM estimates are obtained for each exam administration. Table 6.4 presents the SEM estimates for each ECCE administration. It shows that the SEM estimates are generally stable. Additionally, the SEM estimates as a proportion of the 1000-point scale are very small.

6.5. Trends in Subtest Correlations

Language proficiency measures are typically indirect measures of the trait of language proficiency. Even a direct measure such as a writing task is an indirect measure of the processes involved in composing, in selecting appropriate grammatical constructions, and of the vocabulary resources to which a test taker has access. Language proficiency, therefore, has many facets. For the last thirty years or so, the predominant model of language proficiency has been *communicative language ability* (cf. Bachman, 1990: ch. 4). This characterizes language competence as a multifaceted network of "knowledges" including vocabulary, morpho-syntax, rhetorical organization, conversational rules, language functions, sensitivity to register, and figures of speech.

The ECCE captures evidence of a test taker's communicative language ability at the B2 level of the CEFR through a variety of tasks in the four language skills of listening, reading, writing, and speaking. Section 3.4 described the skills and abilities expect for each language skill. Even though performance on the ECCE is expressed as a pass or fail—that is, a test taker has to pass the ECCE in order to be awarded a certificate—test takers are also issued a score report that presents their results for each test section as a band score. Reporting scores in this way is justifiable if each section can be seen to contribute differentially to the overall ECCE result. Table 6.5 presents the subtest correlations (Spearman's rho) for each year. The correlations range between 0.4 and 0.8, indicating a moderate to strong relationship between the subtests. This is expected since each subtest is intended to measure language proficiency from a different perspective.

Table 6.5 Subtest Correlations (ρ)[1]

Year		Listening	GVR	Writing
2010	GVR	0.732	-	-
	Writing	0.405	0.535	-
	Speaking	0.503	0.560	0.407
2011	GVR	0.758	-	-
	Writing	0.413	0.532	-
	Speaking	0.518	0.571	0.398
2012	GVR	0.762	-	-
	Writing	0.398	0.512	-
	Speaking	0.514	0.567	0.385
2013	GVR	0.786	-	-
	Writing	0.429	0.527	-
	Speaking	0.525	0.576	0.395

1 Correlations are significant at the 0.01 level (2-tailed).

7. Additional ECCE Validity Evidence

Sections 2.2 and 3.4 presented a proposed interpretation of a test taker's ECCE score. The safety of this proposed interpretation is dependent upon the evidence to support it. Test validation is the process of building and augmenting that evidence so that an argument can be presented for the use and interpretation of test scores. Anastasi (1986: 4) and Cronbach (1988) state that the process of gathering validity evidence begins with the design of the test and is never complete. Consequently, validation entails an ongoing research program. Table 7.1 presents proposed claims about the ECCE along with the research evidence available for these claims.

7.1. The different item types and tasks are appropriate for measuring language proficiency at the B2 level on the CEFR

Johnson (2006a, 2006b and 2008) compared the relative difficulty of items on the Examination for the Certificate of Proficiency in English (ECPE) and the ECCE. The ECCE and ECPE are often referred to as sister examinations, aimed at two distinct levels on the CEFR. The ECCE aims at the B2 level while the ECPE aims at the C2 level. The aim of Johnson's work was to establish whether items on the two exams tested at two different levels of language proficiency. Johnson (2006a) and (2006b) used common-person equating to look separately at listening and grammar items for the same population. The linking group (N = 89) took the November–December 2005 ECCE examination as well as the 2005–2006 ECPE examination. To this dataset, Johnson added a group (N = 1111) that took only the November–December 2005 ECCE and another group (N = 2394) that took only the 2005–2006 ECPE. The sample selection was controlled for language background in order to eliminate the possible effect of language background upon performance (particularly on grammar items). Johnson (2006a) found that the linking group generally performed better on the ECCE listening items than on the ECPE listening items, offering preliminary confirmation that listening items on the ECCE and the ECPE are at two different levels of difficulty. This was confirmed by the analysis of item difficulty, which showed a hierarchy of difficulty from the easiest ECCE listening item type to the most

Table 7.1 Proposed Validity Claims about the ECCE and the Research Evidence Available

Proposed Claim	Evidence Available
The different item types and tasks are appropriate for measuring language proficiency at the B2 level on the CEFR	• Johnson, J. S. (2006a) *The relative difficulty of ECCE and ECPE listening section items*, UMELIRR2006–4, University of Michigan. • Johnson, J. S. (2006b) *The relative difficulty of ECCE (05ND) and ECPE (0506) grammar items*, UMELIRR2006–11, University of Michigan. • Johnson, J. S. (2008) *Cross-test item difficulty comparison: ECCE and ECPE listening and reading*, UMELIRR2008–04, University of Michigan
The structure of the test is consistent with its stated construct and with the way in which scores are reported.	• Liao, Y. F. (2007) *Investigating the Construct Validity of the Grammar and Vocabulary Section and the Listening Section of the ECCE: Lexico-Grammatical Ability as a Predictor of L2 Listening Ability*, CaMLA Working Papers, 2007–5.
The language elicited by the speaking and writing sections of the test reflects the domain and/or level of language expected.	• Iwashita, N. & McNamara, T. (2003). *Task and interviewer factors in assessments of spoken interaction in a second language*, Internal Report: University of Michigan • Yang, L. (2005) *A Validation Study of the ECCE NNS and NS Examiners' Conversational Styles from a Discourse Analytic Perspective*, CaMLA Working Papers, 2005–3 • Matice, M. & Briggs, S. (2005) *Task Factors and Their Impact on Spoken Interaction in ECCE Speaking Tests*, UMELIRR2005-1, University of Michigan

difficult ECPE listening item type. Also, on average, the most difficult ECCE listening items were easier than the easiest ECPE listening items.

The results for the grammar items (Johnson, 2006b) were less satisfactory. As in the case of Johnson (2006a), the linking group generally performed better on the ECCE grammar items than on the ECPE grammar items. This was partially confirmed by the analysis of grammar item difficulty in that the mean difficulty logit value for the ECCE grammar items was -0.405 while the mean difficulty logit value for the ECPE grammar items was 0.337. However, the difference between the mean difficulty values was less than one standard deviation, suggesting overlap in item difficulties between the two tests and that some items on each test were incorrectly targeted.

Johnson (2008) replicated these studies with a slightly broader focus, this time looking at the whole reading section (which includes grammar, vocabulary, and reading items) for each test as well as the listening section. Two groups of test takers took the same ECCE examination (May–June 2004) and an ECPE examination (either the 2005–2006 or the 2006–2007 examinations). The N-size for the analysis was relatively small, and one of the subgroups was clearly at a higher proficiency level, presenting problems for the equating design that was used. Consequently, results should be interpreted with caution. Nevertheless, the data analysis indicated that, as in the 2006 studies, the ECPE items were generally more difficult than the ECCE items. As in the earlier studies, there were overlaps in difficulty at both ends of the difficulty spectrum, suggesting that some ECCE and ECPE items were mistargeted.

These studies provided some insight into whether the ECCE listening and reading section items and tasks appropriately target the level of the test. Since this time, however, both sections have undergone revisions and new item types have been introduced. In order to confirm the success of these revisions, it would be useful to conduct a similar investigation, perhaps using common-person equating in order to avoid some of the problems that arose with the Johnson (2008) study. It is also important to note that, while these studies provide insight into the relative difficulty of items on the ECCE and ECPE, they do not provide explicit evidence that the ECCE items are measuring at the B2 level on the CEFR. A standard-setting study that explicitly investigates this question has been prioritized.

7.2. The structure of the test is consistent with its stated construct and the way in which scores are reported.

Liao (2007) has investigated the construct validity of the listening section and the grammar and vocabulary items of the ECCE. Liao first explored the underlying structure of the listening, grammar, and vocabulary items, and then examined to what extent lexico-grammatical knowledge predicted L2 listening ability. In doing so she aimed to: (1) identify the factor structure of the listening section, (2) identify the factor structure of the grammar and vocabulary subsection, and (3) investigate the relationship between lexico-grammatical knowledge and listening ability.

Liao's study used data from the listening, grammar, and vocabulary sections from the 2003 administration of the ECCE. Before any analysis was conducted, the exam items were coded by four trained and experienced ESL teachers to determine what aspect of language they measured. The grammar and vocabulary (GV) items were coded using Purpura's (2004) model of grammatical knowledge, which proposed two dimensions of grammatical knowledge, either literal meaning (i.e. lexical meaning) or grammatical form (i.e. lexical, morphosyntactic, or cohesive form). The listening items were coded using Buck's (2001) and Wagner's (2002, 2004) theoretical models of listening ability, which divided listening ability into two traits, the ability to listen for explicitly stated information, and the ability to listen for implicitly stated information.

In order to examine the underlying traits of the GV items and listening items, exploratory factor analysis (EFA) was performed for both. These analyses provided support for the hypothesis that the listening section consisted of two factors (explicit and implicit) and that the GV section also measured two factors (form and meaning). The study then used item level structural equation modeling (SEM) to learn more about the factor structure of the listening and GV sections. The study examined correlated two factor models for the GV and listening items, using the item coding to determine which items would load on which factor. It was found that both models fit the data well, and therefore provided evidence that the two factor solutions provided reasonable explanations of the correlations between the observed variables.

In addition to this analysis, a series of SEMs were performed to investigate the relationship between lexico-grammatical knowledge and listening ability. They showed that grammatical knowledge was a

moderately strong predictor of listening ability. Both the form and meaning factors were strong predictors of listening ability, particularly the ability to listen for implicit information. The evidence of a close positive relationship between lexico-grammatical knowledge and L2 listening test performance found in this study is in accordance with the findings of Mecartty's (2000) study. Discriminant analysis was also performed to examine the extent to which the predictor variables (form and meaning) correctly classified the test takers L2 listening comprehension ability. The results showed that using both of these factors as predictors resulted in a moderately high percentage of correctly classified test takers.

Overall, this study provides validity evidence that the traits measured by the items in this study are consistent with the stated construct of the ECCE. The listening section was found to measure two traits, listening for explicit information and listening for implicit information, which conforms to the models of L2 listening ability posited by Buck (2001) and Wagner (2004). Similarly, the grammar and vocabulary items were found to measure two traits, grammatical form and literal meaning, which conforms to the theoretical model of grammatical knowledge presented in Purpura's (2004) work.

7.3. The language elicited by the speaking and writing sections of the test reflects the domain and/or level of language expected.

The ECCE speaking test has been analyzed in three different studies. The first, by Iwashita and McNamara (2003), examined the validity of the ECCE speaking section by examining the three main components of the assessment: the rating scale, the interviewer, and the tasks. The study addressed several questions related to these aspects: (1) to what extent do quantitative scores represent qualitatively different performances, (2) how do interviewer techniques vary among interviewers, (3) what stylistic variations among interviewers have an impact on test-taker performance, (4) do different tasks elicit quantitatively and qualitatively different performances, and (5) what features, if any, characterize the discourse of interlocutors and test-takers on these tasks.

The data analyzed in this study comprised a set of videotaped interviews that were conducted expressly for research purposes. Sixteen test takers took two interviews (the prompt and interviewer were different for both interviews) so the dataset contained 32 total assessments. The recorded performances were also transcribed, so that discourse analysis could be performed. Three experienced raters evaluated each taped performance using the ECCE rating scales. The test takers were assigned one holistic score and seven analytic scores for their performance. The study then analyzed these performances and performed discourse analysis, which focused on three primary features: fluency/intelligibility (delivery and articulation scores), grammar/vocabulary (morphology and vocabulary scores), and functional language use (elaboration and initiative scores). Several aspects of test taker performance (such as pause time, filled pauses, and unfilled pauses for fluency/intelligibility) were identified as suitable measures of these feature based on the literature.

During the analysis of the rating scale, Iwashita and McNamara (2003) found that there were notable qualitative and quantitative differences between high and low scoring test takers. Specifically, those who obtained higher holistic scores had less total unfilled pause time, fewer total unfilled pauses, more vocabulary knowledge, more elaboration, and were more likely to initiate interactions with the interviewer. They also found that the test takers performance differed greatly, both in terms of the quality and quantity of language, between the first and second interviews. Iwashita and McNamara (2003) speculated that these differences could be accounted for by differences in the interview occasion (either first or second), the examinee level (the second interview scored higher than the first), and the interviewers (different interviewers conducted the two interviews).

When the performance of the interviewers (and the different styles and techniques they employed) was analyzed, the authors found that there was considerable variation between interviewers, particularly in how they solicited responses from test takers. Each interviewer appeared to have their own distinctive style when conducting the exam. These variations were most evident in how the interviewers transitioned from one task to another, and in the way they provided instructions for the tasks. While the interviewers were generally internally consistent in their administration of the exam, some differences were observed based on examinee proficiency level. For lower scoring test takers examiners tended to be more dominant in their interactions than they were with the other examinees.

Analysis of the effects the different tasks had on test taker performance revealed that there was little difference in the quality of a test taker's performance between the tasks. Iwashita and McNamara (2003) speculated that these findings may be dependent on the way in which all the tasks were treated by the interviewers and examinees. In this dataset, all of the tasks were generally treated as personal with the examiners asking several questions relating to the examinee's personal experiences for all tasks. While this was desired during task 1, the other tasks were meant to be approached differently. The only notable difference found between tasks was that low scoring test takers struggled with tasks 2 and 3, and appeared to not fully understand what they were supposed to do or what questions they were expected to ask.

In a small scale study, Matice and Briggs (2005) extended the work done by Iwashita and McNamara (2003) following the suggestions provided for future research. Their study had three main purposes: (1) to investigate the discourse produced in ECCE speaking tests using two different prompt types, (2) to evaluate recent changes made to the tasks, to obtain information that can guide the development of future prompts, and to evaluate the instructions given to the examiners about how to conduct the test, and (3) to confirm the findings of Iwashita and McNamara (2003) regarding the features of the scale that most impact holistic scores.

This study analyzed the performance of 7 test takers on two different forms of the ECCE speaking test. In total, there were 14 recorded and transcribed examinations, 7 using a traditional prompt and 7 using a problem solving type prompt. The order the exam forms were administered was alternated so that the effect of form order on the analysis would be minimized. Additionally, all of the test takers were female and took the exam in either Greece or Uruguay. This study analyzed the discourse produced by both test takers and examiners in ECCE speaking tests. For test takers, the features examined included the quantity of words, number of pauses and their length, length of each response, grammar and miscommunication on the longest responses, and vocabulary. Examiners were studied for features such as the types of questions asked, the use of extension questions, the amount of examiner speech, and the transition between tasks.

While the data sample used in this study was very small and not representative of a typical ECCE administration, the authors felt that the results of the study permitted them to draw several conclusions and make recommendations for future ECCE speaking tests. The study found that the errors in grammatical control for the higher level test takers seemed more related to morphology than structure, and the vocabulary limitations of lower level test takers was related to both receptive vocabulary and word retrieval. Several of the revisions to the ECCE speaking test, such as instructing examiners to use elaboration questions, instructing test takers to use information provided by the examiner in task 2, and instructing examiners to ask test takers why they did not choose another option, were found to be effective. The study also found that while the two prompt forms produced a similar amount of content, the prompts had an impact on fluency and average utterance length. Matice and Briggs (2005) concluded that elaboration questions for problem solving prompts should be less personal, while elaboration questions for traditional prompts should try and relate the subject to the test takers personal experience or background. Finally, the study made several recommendations to improve the rating scale. Matice and Briggs (2005) suggested that the scale descriptors for grammatical control be revised with regard to both structure and morphology. They also suggested that revisions be made to accommodate low level test takers who showed elaboration and initiative. Finally, they recommended that the extent to which utterance length leads to miscommunication should be incorporated into the scale.

In another study of the ECCE speaking exam, Yang (2005) examined the conversational styles of native and nonnative speaking (NS and NNS) examiners to observe their effect on the assessment of the test takers oral proficiency. Specifically, this study aimed to: (1) see if there were overall differences between the amount and types of eliciting and non-eliciting moves in discourse produced by the NS and NNS examiners, (2) identify non-eliciting discourse features that do not encourage examinees to elaborate their replies, (3) identify non-eliciting discourse features that do not encourage the test takers to seek information, and (4) discover any differences between NNS and NS examiners in the amount and types of non-eliciting discourse features.

Yang used twenty live recordings of ECCE speaking tests (nine by NNS examiners, eleven by NS examiners) administered in 2004. These dialogues were transcribed and then analyzed using a task specific discourse analysis (DA) model for analyzing spoken discourse. This model was based on an overall model for analyzing an exchange in the interactive discourse in oral proficiency tests,

and modified, using the general guidelines from the examiner's manual, for the specific tasks present on the ECCE.

The DA resulted in a large list of discourse features that either did not encourage test takers to elaborate their replies, or did not encourage them to seek information (i.e. agreeing, back-channeling, correcting mistakes, etc.). The study found that while the ECCE Speaking Test examiners generally followed the developers' guidelines for eliciting test taker language, there were deviations from the requirements by both the NNS and NS examiners. Overall, the NNS examiners were less likely to encourage test takers to elaborate on their responses. Compared with the NS examiners, the NNS examiners performed less eliciting behavior, and more non-eliciting behavior. The amount of eliciting discourse features used by the NNS examiners was sometimes half of that used by NS examiners, while the amount of non-eliciting features made by NNS examiners was often twice that of NS examiners.

Taken together, these three studies highlighted several issues in the consistency of the administration of the ECCE speaking test. These results informed revisions to the design of the test and the examiner training materials. The exam is now semi-scripted in order to ensure that test takers receive very similar test taking experiences regardless of their speaking test examiner. Ongoing monitoring of ECCE speaking tests suggests that these changes (implemented in 2008) have resolved the issues presented, and have resulted in a more consistent and higher quality speaking test that better reflects the test construct. Nevertheless, it would be useful to conduct a follow-up study to confirm this.

7.4. Future Research Needed

The research already completed has begun the work of building a validity argument for the ECCE. However, there are still many avenues to be pursued. Proposals would be welcomed for further research, particularly work that could support the following claims about the ECCE:

- The content of the test is representative of the kinds of oral and written texts and tasks that might be encountered by high intermediate learners of English at the B2 level on the CEFR.
- The writing and speaking rating scales reflect the features of language proficiency expected of learners of English at the B2 level on the CEFR.
- The language processes and linguistic knowledge that the test takers use to successfully complete the ECCE reflects the language knowledge and processes expected at the B2 level on the CEFR.
- Performance on the ECCE is related to other indicators of language proficiency.
- ECCE test results are used appropriately
- The ECCE has positive consequences for stakeholders.

8. References

Anastasi, A. (1986). Evolving concepts of test validation, *Annual Review of Psychology*, 37, 1–15.

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

Buck, G. (2001). *Assessing listening*. Cambridge, UK: Cambridge University Press.

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

Cronbach, L. J. (1988). Five perspectives on the validity argument, in H. Wainer and H.I. Braun (Eds.) *Test Validity* (pp. 3–18), Hillsdale, N.J.: Lawrence Erlbaum Associates.

Field, A. (2005). *Discovering statistics using* SPSS, London: Sage Pu8blications Inc.

Iwashita, N. & McNamara, T. (2003). *Task and interviewer factors in assessments of spoken interaction in a second language*, Internal Report: University of Michigan

Johnson, J. S. (2006a) *The relative difficulty of ECCE and ECPE listening section items*, UMELIRR2006–4, University of Michigan.

Johnson, J. S. (2006b) *The relative difficulty of ECCE (05ND) and ECPE (0506) grammar items*, UMELIRR2006–11, University of Michigan.

Johnson, J. S. (2008) *Cross-test item difficulty comparison: ECCE and ECPE listening and reading*, UMELIRR2008–04, University of Michigan.

Leckie, G. & Baird, J. (2011) Rater effects on essay scoring: a multilevel analysis of severity drift, central tendency, and rater experience, *Journal of Educational Measurement*, 48(4): 399–418.

Liao, Y. F. (2007). *Investigating the Construct Validity of the Grammar and Vocabulary Section and the Listening Section of the ECCE: Lexico-Grammatical Ability as a Predictor of L2 Listening Ability,* CaMLA Working Papers, 2007–5.

Matice, M. & Briggs, S. (2005) *Task Factors and Their Impact on Spoken Interaction in ECCE Speaking Tests*, UMELIRR2005-1, University of Michigan.

Mecartty, F. H. (2000). Lexical and grammatical knowledge in reading and listening comprehension by foreign language learns of Spanish. *Applied Language Learning*, 11(2), 323–348.

Purpura, J. E. (2004). *Assessing grammar*. Cambridge, UK: Cambridge University Press.

Wagner, E. (2004). *A Construct Validation Study of the Extended Listening Sections of the ECPE and MELAB*, CaMLA Working Papers, 2004–2.

Yang, L. (2005). *A Validation Study of the ECCE NNS and NS Examiners' Conversational Styles from a Discourse Analytic Perspective*, CaMLA Working Papers, 2005–3.