2009–2010
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
www.CambridgeMichigan.org

03/2013

Table of Contents

List of Tables

1. Introduction

The Examination for the Certificate of Proficiency in English (ECPE) is a test of general language proficiency for advanced learners of English. It is administered twice annually in CaMLA test centers worldwide. During 2009 and 2010 the exam was administered four times, twice each year.

This report provides test users with technical information about the ECPE. Part 2 provides general information about the test and a proposed interpretation of ECPE test scores. In Part 3, the report explains how each section is scored and equated, and the procedures for reporting scores. It also gives guidelines for score use in decision-making. Part 4 describes the changes to the ECPE during 2009 and 2010, including the introduction of a new speaking test. Subsequent parts of the report focus on statistical analyses of test data from the four administrations in 2009 and 2010. Part 5 discusses the ECPE test-taking population, looking particularly at the distribution of test takers by gender and by age. Part 6 looks at trends in the ECPE test results by section, age, and gender. It examines trends in reliability estimates, standard error of measurement, and rater agreement statistics. It also presents subtest correlations for each administration in 2009 and 2010. The final section of the report reviews the validity evidence currently available to support CaMLA's proposed interpretation of the ECPE certificate.

2. Description of the ECPE

2.1. General Description

The ECPE is a standardized advanced-level English as a foreign language (EFL) examination. The ECPE is a test of general language proficiency in a variety of contexts; it assesses the four component skills of listening, reading, writing, and speaking through a combination of tasks.

The ECPE is aimed at the C2 level of the Common European Framework of Reference (CEFR) and is valid for the lifetime of the recipient. It is recognized in several countries as official documentary evidence of advanced proficiency in the English language and can be used for academic and professional purposes. It is accepted by some universities as evidence of proficiency in English if the certificate has been received within the past two years.

CaMLA is committed to excellence in 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 securely administered in a way that is fair and accessible to test takers and that the ECPE 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 ECPE is aimed at the C2 level of the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (Council of Europe, 2001) (CEFR). Language users at this proficiency level:

> Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express [themselves] spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.
>
> (Council of Europe, 2001: 24)

Therefore, ECPE certificate holders are expected to be comfortable engaging with abstract ideas and concepts. They are interactive oral English speakers; they contribute to the development of a discussion, can

understand conversational questions, can grasp both the gist and details of a conversation delivered in Standard American English, and can understand extended spoken discourse. They should also have a wide-ranging and flexible vocabulary as well as a sound grasp of English grammar. They can understand written materials that are encountered in both general and specialized professional contexts as well as in university-level reading. Additionally, they are able to communicate in standard written English with good expression and accuracy.

2.3. Test Structure

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

www.CambridgeMichigan.org/exams/ecpe/format

Table 2.3 Format and Content of the ECPE

Section	Time	Description	Number Of Items
Speaking	30–35 minutes	Test takers participate in a semistructured, multistage task involving two test takers and two examiners.	1 task
Writing	30 minutes	Test takers write an essay based upon one of two topic choices.	1 task
Listening	35–40 minutes	**Part 1** (multiple choice) A short recorded conversation is accompanied by three printed statements. Test takers choose the statement that conveys the same meaning as what was heard, or that is true based upon the conversation.	50
		Part 2 (multiple choice) A recorded question is accompanied by three printed responses. Test takers choose the appropriate response to the question.	
		Part 3 (multiple choice) Three recorded talks, such as those that might be heard on the radio, are each followed by recorded comprehension questions. The questions and the answer choices are printed in the test booklet. Test takers choose the correct answer from the choices.	
Grammar Cloze Vocabulary Reading	75 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.	40
		Cloze (multiple choice) After reading a passage from which words have been removed, test takers must choose one of four words that best fills a missing word slot in terms of grammar and meaning.	20
		Vocabulary (multiple choice) An incomplete sentence is followed by a choice of words to complete it. Test takers must choose the option that best completes the sentence in terms of meaning.	40
		Reading (multiple choice) Four reading passages are followed by comprehension questions. Test takers choose the correct answer from the printed answer choices.	20

3. Scoring and Reporting of ECPE Results

3.1. Explanation of Scoring for Each Section

The speaking and writing sections are graded according to scales established by CaMLA (see the CaMLA website for speaking and writing rating scales). The speaking section is conducted and assessed by two certified oral examiners. The writing section ratings are assigned by specialized raters trained and certified according to CaMLA standards. All essays are scored by at least two raters.

The listening and grammar, cloze, vocabulary, and reading (GCVR) sections of the ECPE are computer-scored by CaMLA. Each correct answer carries equal weight within each section and there are no points deducted for wrong answers. A scaled score 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.

If a test taker's scores (the rating for speaking and writing; the scaled score for listening and GCVR) meet the cutoff level in a section, they are given a pass for that section of the exam. Test takers who pass three sections with a Low Pass (or higher) and receive no less than a Borderline Fail in one section will be awarded an ECPE certificate. Those with Honors scores on all four sections are awarded a Certificate of Proficiency with Honors.

3.2. Equating Procedures

The approach adopted is common item equating using item difficulties. The procedure is as follows:

- Item difficulties from previous administrations are stored in a database that includes placement of linking items in the current administration.
- A scatterplot of item difficulties is generated with previous equated values on the X axis and current estimated values on the Y axis.
- $\theta^{eq} = \alpha\theta + \beta$ where α and β location and scale factors are calculated by fitting a SD line to the scatterplot (above)—the SD line gives an invertible transformation

Scale and location factors are computed separately for the listening and GCVR sections and implemented in a scoring run in BILOG-MG.

3.3. Procedures for Reporting Scores

All test takers will receive an Examination Report that shows their overall performance as well as the levels for each test section (see the appendix for a sample report). 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 ECPE (Honors/Pass/Fail)
- Section results with a brief description of the test taker's performance.

ECPE section scores are reported in five bands. These levels of performance, from highest to lowest, are:

- Honors (HP)
- Pass (P)
- Low Pass (LP)
- Borderline Fail (BF)
- Fail (F)

3.4. Interpretation of Scores for Each Section

As stated in the description of the exam (section 2.2), the ECPE is aimed at the C2 (Mastery) 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 well-structured presentations and take part in a wide variety of conversations, using formal and informal language. They can structure their utterances logically and can use grammatical structures and vocabulary flexibly in order to convey their precise meaning. Where necessary, they can reformulate in order to help their listener to understand.

Writing They are able to communicate their ideas fully in clear, smoothly flowing language. They can structure their text logically to present an effective argument and can use grammatical structures and vocabulary flexibly in order to convey their precise meaning.

Listening They are able to understand conversations, debates, and monologues on a

wide range of topics, including topics that are abstract and complex. They are comfortable listening to speech delivered at a native-speaker rate, in both formal and informal contexts. As long as there are opportunities to seek clarification, they can understand a wide range of idiomatic expressions and colloquialisms.

Reading They are able to understand written materials on a wide range of topics, including topics that are abstract and complex. They are able to identify the main idea of a text, and can read quickly to locate important details as well as infer attitudes or make connections between ideas.

Use of English They are able to maintain a high degree of grammatical accuracy even under time pressure, such as in circumstances where they have limited opportunity to plan their speech or writing. They can use grammatical structures to give emphasis. Similarly, their vocabulary is wide-ranging and flexible, allowing them to convey fine shades of meaning and avoid ambiguity.

3.5. Guidelines for Decision-Making

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

4. Changes to the ECPE in 2009 and 2010

Four major changes to the ECPE have been introduced in the period covered by this report. Two of these changes were scoring changes and two were changes to the design of the test.

4.1. Changes to the Design of the Test

Printing question stems in the test booklets for the listening interview item type (introduced in November 2009)

Until November 2009, the question stems for Part 3 of the listening section (the radio reports) were presented in audio format only. As part of its ongoing program of test review and renewal, and in order to provide test takers with a clearer purpose while listening, CaMLA made a small change to the presentation of the question stems for Part 3 of the listening section; they are now also presented in written form in the test booklets. This allows test takers to read the questions while they are listening to the radio report.

New ECPE speaking test (introduced in May 2009)

The new ECPE speaking test was developed over a three-and-a-half-year period beginning in September 2005. The project included a full construct review, beta-testing of test designs and task types, as well as the development of an extensive examiner training package.

The new ECPE speaking test allows test takers to demonstrate the full range of their speaking ability while performing a multistage, semistructured, decision-making task. The task might involve selecting a candidate for a particular job, or selecting an advertising campaign (see www.CambridgeMichigan.org/ecpe for more details). Working with one or two partners, test takers are each given descriptions of two different options. Test takers collaborate to learn about all the options available, decide on a single option, and then present and defend that option.

4.2. Scoring Changes

Five-point rating scale for the writing section (introduced in May 2009)

Up until May 2009, the rating scale for the writing section comprised four bands, A–D, where D was the only failing band. In the final quarter of 2008, the rating scale was expanded to a five-point scale with two failing bands, D and E. The new scale was tested in November 2008 before being finalized. In conjunction with the new writing rating scale, CaMLA revised and expanded the writing section rater training and certification materials that are provided for ECPE essay raters.

Modification of the requirements for passing the ECPE (introduced in May 2009)

Up until May 2009, test takers were required to pass all four sections of the ECPE before they could be awarded an ECPE certificate. However, the major revisions to the ECPE speaking test (including a new, detailed five-point rating scale) and the introduction of a revised, five-point rating scale for the ECPE writing section harmonized the assessment for all the sections of the ECPE and allowed more precise measurements of a test taker's performance.

Consequently, the requirements for passing the ECPE overall were changed to allow for one borderline fail. Test takers who receive a minimum of a Low Pass in three sections of the ECPE and no lower than a Borderline Fail in the fourth section are awarded an ECPE certificate.

5. ECPE Test-Taking Population

This section presents an overview of the test takers who took the ECPE during the period covered by this report, providing information about the countries in which the exam was administered, as well as the distributions for test taker gender and age.

5.1. Test-Taker Numbers

Table 5.1 presents the number of test takers who attempted the ECPE in 2009 and 2010. It shows that the number of test takers who take the ECPE appears to be holding relatively steady at around 40,000 test takers per year.

Table 5.1 Number of Test Takers who Attempted the ECPE in 2009 and 2010[1]

Administration	Number of Test Takers
MJ2009	25,976
N2009	14,893
MJ2010	25,606
N2010	15,341

5.2 Range of Countries in Which the Test is Administered

Table 5.2 shows that the ECPE was administered in 21 countries during 2009 and 2010. The coverage of test centers is best in Europe and Latin America. CaMLA would like to increase the number of test centers in other parts of the world, including Asia.

Table 5.2 Countries Where the ECPE was Administered in 2009 and 2010

Argentina	El Salvador	Mexico
Austria	Ecuador	Paraguay
Belgium	Greece	Peru
Bolivia	Guatemala	Portugal
Brazil	Italy	Spain
Bulgaria	Jordan	Uruguay
Colombia	Lebanon	Vietnam

1 The totals presented here reflect the total candidature for each testing session and include examinees whose performances were received late. As a result, these numbers may differ slightly from the numbers presented in the Administration Reports for each testing session.

5.3. Gender Distribution

Every ECPE test taker completes a registration form that asks for their gender. Cases where information is not given or is not correctly given are treated as missing data. Table 5.3 shows the distribution of test takers by gender. The data shows that just under two-thirds of the ECPE candidature tends to be female. This distribution is similar for each administration of the exam. This could be because, in our major market, the primary use of the ECPE qualification is as a teaching qualification, and because the majority of English language teachers in this market are female.

Table 5.3 Distribution (in %) of Male and Female Test Takers in 2009 and 2010

	Male (%)	Female (%)	Missing Data (%)
MJ2009	39.18	60.75	0.05
N2009	38.34	61.93	0.27
MJ2010	39.11	60.82	0.06
N2010	36.98	62.75	0.25

5.4 Age Distribution

The ECPE registration form also asks for the test takers' date of birth. Cases where information is not given or is not correctly given are treated as missing data. Table 5.4 (on the next page) shows the distribution of test takers by age. The data shows that the majority of ECPE test takers are less than 20 years old, with a sizeable proportion in the 13–16 age group. It indicates that the majority of the ECPE candidature take the ECPE while still in formal schooling and before they attend university.

Table 5.4 Age Distribution (in %) for 2009 and 2010

Age	MJ2009 (%)	N2009 (%)	MJ2010 (%)	N2010 (%)
≤ 12	0.08	0.13	0.12	0.17
13–16	43.53	47.99	46.15	51.33
17–19	12.84	10.90	12.09	10.06
20–22	19.40	16.92	17.40	15.38
23–25	10.48	9.24	9.67	8.67
26–29	6.60	6.58	7.00	6.76
30–39	5.07	5.82	5.54	5.52
≥ 40	1.90	2.25	1.99	2.07
Missing Data	0.06	0.17	0.04	0.03

Looking more closely at the largest test-taking group, the 13–16 age range, table 5.4 also shows a discernible rise year-on-year. The candidature in this age range has risen across the whole period—that is, between the May 2009 (MJ2009) ECPE administration and the November 2010 (N2010) administration. It has also risen between the two May administrations (MJ2009 and MJ2010) and the two November administrations (N2009 and N2010). This suggests a trend towards taking the ECPE at an ever-earlier age.

6. ECPE Results and Test Statistics

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

Table 6.1.1 shows the distribution (in %) of ECPE passes for the 2009 and 2010 administrations.

Table 6.1.1 Distribution (in %) of ECPE Passes for the 2009 and 2010 Administrations

	Honors Pass	Pass	Fail
MJ2009	0.17	56.40	43.30
N2009	0.30	59.10	40.60
MJ2010	0.16	60.18	39.66
N2010	0.25	61.53	38.22

The table shows a discernible upwards trend in the pass rate; since the MJ2009 administration, there has been a 5.13 percentage point rise in the pass rate. Because the ECPE is carefully monitored to ensure that each test form is aimed at the C2 level on 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 best explored by analyzing the performance trends for each section of the exam.

The upwards trend in the overall pass rate for the ECPE is addressed in more detail as part of comparison of the score distribution and pass rates for the individual sections.

Listening

Table 6.1.2 shows the distribution (in %) of passes on the listening section for the 2009 and 2010 administrations.

There appears to be an upwards trend in the % of passes on the listening section. 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 ECPE equating procedure, suggests that this is a plausible explanation. Test takers for the 2009 and 2010 exams have performed better on the common items than previous test populations.

GCVR

Table 6.1.3 shows the distribution (in %) of passes on the GCVR section for the 2009 and 2010 administrations.

As with the listening section, there appears to be an upwards trend (albeit less marked) in the % of passes on the GCVR section. Like the listening section, one explanation could be that the candidature is becoming more able. Performance on the GCVR common items suggests that this is a plausible conclusion.

Table 6.1.2 Distribution (in %) of Passes on the Listening Section for the 2009 and 2010 Administrations

	Honors Pass	Pass	Low Pass	Borderline Fail	Fail	Total Pass
MJ2009	9.40	21.90	28.50	11.50	28.40	60.10
N2009	14.40	22.40	25.50	10.50	27.30	62.20
MJ2010	9.82	25.94	31.37	11.83	21.79	66.38
N2010	5.85	23.70	34.45	13.83	22.12	64.05

Table 6.1.3 Distribution (in %) of Passes on the GCVR Section for the 2009 and 2010 Administrations

	Honors Pass	Pass	Low Pass	Borderline Fail	Fail	Total Pass
MJ2009	6.60	23.20	35.10	13.70	21.20	65.10
N2009	9.20	25.90	34.10	11.70	19.10	69.10
MJ2010	11.21	24.01	31.21	11.92	21.65	66.43
N2010	11.02	24.58	32.19	11.94	20.27	67.79

Writing

Prior to May 2009, the ECPE writing section was rated on a four-point scale and the overall pass rate for the section tended to be 90% or higher. This was of considerable concern since it indicated that the writing section of the ECPE was out of line with the listening and GCVR sections.

Table 6.1.4 shows the distribution (in %) of passes on the writing section for the 2009 and 2010 administrations. It shows the pass rate for the writing section after the introduction of the five-point rating scale and the implementation of revised rater training and qualification materials.

As the table indicates, unlike the listening and GCVR sections, the pass rate for the writing section is trending downwards. Assuming that the % pass rate on all sections of the ECPE should be approximately the same, this is a good sign.

However, the % pass rate for the writing section is still considerably higher than that for the listening and GCVR sections. It also appears that lowest (fail) band is relatively underutilized for the writing section. While 20% or more of the scores for the listening and GCVR sections fall into the Fail category, less than 1% of the scores for the writing section fall into this band. Because the overall score for the ECPE allows for a single borderline fail, underuse of the Fail category in the writing section might be contributing to the upwards trend in the pass rate for the ECPE overall.

It is as important to note that the highest (honors pass) band is underutilized for this section. While 6–15% of the scores for the listening and GCVR sections fall into the Honors Pass category, only 0.46–1% of the scores for the writing section fall into this band. Additionally, more than two-thirds of the test-taking population is awarded a Low Pass (the minimal passing grade). This pattern can partially be explained by the fact that test takers will register for the ECPE 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 and/or the design of the rating system. Reviews of rater training materials and the design of the rating system will need to pay particular attention to the use of the Borderline Fail and Fail categories.

Speaking

Prior to May 2009, the ECPE speaking test was an interview conducted by an examiner with a single test taker and was rated on a four-point scale. The overall pass rate for this section was regularly at approximately 98%. As in the case of the writing section, this was a source of concern since it indicated that the speaking section of the ECPE was out of line with the listening and GCVR sections.

Table 6.1.4 Distribution (in %) of Passes on the Writing Section for the 2009 and 2010 Administrations

	Honors Pass	Pass	Low Pass	Borderline Fail	Fail	Total Pass
MJ2009	0.56	13.30	70.70	14.70	0.51	84.79
N2009	1.00	15.40	73.20	10.50	0.40	89.10
MJ2010	0.46	9.29	66.20	23.42	0.63	75.95
N2010	0.70	11.84	69.44	17.65	0.37	81.98

Table 6.1.5 Distribution (in %) of Passes on the Speaking Section for the 2009 and 2010 Administrations

	Honors Pass	Pass	Low Pass	Borderline Fail	Fail	Total Pass
MJ2009	4.60	18.50	51.30	23.30	1.87	74.83
N2009	3.80	19.60	55.00	20.00	1.40	78.60
MJ2010	4.55	21.12	56.40	16.68	0.88	82.44
N2010	4.98	20.70	57.47	15.86	0.95	83.19

Table 6.1.5 shows the distribution (in %) of passes on the speaking section for the 2009 and 2010 administrations. It indicates that the pass rate for the speaking section has dropped since the introduction of the new speaking test. However, the % pass rate for the speaking test is still considerably higher than that for the listening and GCVR sections. Additionally, there appears to be an upwards trend in the pass rate since the introduction of the new speaking test in May 2009. This can partly be attributed to increased familiarity with the new test format and more explicit preparation in ECPE preparation classes. However, there are two other notable trends. The first is a drop of almost 8 percentage points in the Borderline Fail category. The second is the generally low (and dropping) % of test takers who are awarded a Fail. Both of these patterns suggest a tendency for examiners to avoid the two fail categories of Borderline Fail and Fail. Because the overall score for the ECPE allows for a single borderline fail, underuse of the Fail category in the speaking test might be contributing to the upwards trend in the pass rate for the ECPE overall.

The performance trends for individual sections of the ECPE suggest that test takers are increasingly well prepared for the examination but also that the score trends for the ECPE writing and speaking sections are out of line with the score trends for the listening and GCVR sections. CaMLA provides regular updates to the training and certification materials and gives feedback to examiners after each test administration. This cycle of training and feedback opportunities is designed to improve the quality of the rating process but is clearly an ongoing task.

6.2. Distribution of Results by Age and Gender

Table 6.2.2 presents the percentage (%) of test takers from each age group who received an overall Pass for the 2009 and 2010 administrations.

Table 6.2.2 Percentage of Test Takers from Each Age Group Who Received an Overall Pass for the 2009 and 2010 Administrations[2]

Age	MJ2009	N2009	MJ2010	N2010
≤ 12	80.00	83.33	66.67	70.00
13–16	61.28	61.64	64.74	63.50
17–19	55.70	62.78	58.24	67.02
20–22	53.13	55.49	57.17	58.26
23–25	54.53	59.29	54.97	55.56
26–29	52.56	62.96	58.20	61.13
30–39	55.90	63.71	63.66	60.68
≥ 40	45.57	55.85	51.73	52.80

It is important to note that the percentage of the test-taking population in the two extreme age groups is very small (see Table 5.4, above), so the results for these two groups should be interpreted with caution. Nevertheless, the data in Table 6.2.2 suggests that younger test takers (i.e., ≤ 19 years old and probably still in full-time, secondary education) tend to perform well on the ECPE, as do test takers in the 26–39 age group. In order to establish whether these differences were meaningful, we ran cross-tabulations and chi-square tests for each administration. Table 6.2.3 summarizes the Pearson Chi-Square value for each administration, 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. The higher the value, the stronger the association.

2 % of missing cases for each administration ranges between 0.03% and 0.21%.

Table 6.2.3 Chi-Square Test Results for the Age*ECPE Pass Analyses of the 2009 and 2010 Administrations

	χ^2	df	*p*	Cramer's *V*
MJ2009	156.58	7	< 0.00	0.078
N2009	44.84	7	< 0.00	0.055
MJ2010	171.10	7	< 0.00	0.082
N2010	73.98	7	< 0.00	0.069

Table 6.2.3 shows that, for each administration, there was a significant association between the age group and whether or not a test taker would pass the examination. However, the Cramer's *V* measure indicates that the effect size for each administration is small; that is, these differences might not be sufficiently large to be meaningful.

On the basis of these analyses, it is not possible to say with any certainty that particular age groups of test takers are more likely to pass the ECPE. However, if the trend towards increasing numbers of younger test takers continues (see 5.4, above), it would be interesting to perform these analyses for subsequent administrations.

Turning now to the distribution of results by gender, Table 6.2.4 presents the percentage (%) of male and female test takers who received an overall Pass for the 2009 and 2010 administrations.

Table 6.2.4 Percentage of Male and Female Test Takers who Received an Overall Pass for the 2009 and 2010 Administrations[3]

Gender	MJ2009	N2009	MJ2010	N2010
Male	61.42	64.29	66.25	68.08
Female	53.57	58.12	56.59	58.05

The data in Table 6.2.4 suggests that male test takers tend to perform better on the ECPE than female test takers. As in the case of the age groups (Table 6.2.3, above), we ran cross-tabulations and chi-square tests for each administration in order to establish whether the pattern was meaningful.

3 % of missing cases for each administration ranges between 0.04% and 0.05%.

Table 6.2.5 Chi-Square Test Results for the Sex*ECPE Pass Analyses of the 2009 and 2010 Administrations

	χ^2	df	*p*	Cramer's *V*
MJ2009	155.25	2	< 0.00	0.077
N2009	60.84	2	< 0.00	0.064
MJ2010	240.69	2	< 0.00	0.097
N2010	152.81	2	< 0.00	0.100

Table 6.2.5 summarizes the Pearson Chi-Square value for each administration, 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 administration, there was a significant association between the sex of the test taker and whether or not they would pass the examination. However, the Cramer's *V* measure indicates that the effect size for each administration is small. On the basis of these analyses, it is not possible to say with any certainty that male test takers are more likely to pass the ECPE than female test takers. Still, given the consistently higher numbers of female test takers (see 5.3, above), it would be important to see if these findings hold true for subsequent administrations.

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, 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, even when the test conditions are carefully controlled. 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 exams such as the ECPE, which aim at a specific proficiency level, a reliability estimate of 0.80 and above is considered acceptable.

Table 6.3.1 presents the reliability estimates for the listening and GCVR sections during 2009 and 2010. 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 120 items. Nevertheless, both sections are consistently above 0.80.

Table 6.3.1 Reliability Estimates for the Listening and GCVR Sections During 2009 and 2010

Section	MJ2009	N2009	MJ2010	N2010
Listening	0.85	0.84	0.86	0.85
GCVR	0.94	0.92	0.93	0.93

In the case of performance tests such as the writing and speaking sections of the ECPE, 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 rater agreement for the writing and speaking sections. The examiners 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. If agreement cannot be reached between two out of three raters, the essay is sent to CaMLA for review and final scoring. It is important to note that *the final score awarded for each ECPE essay 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 an essay should be awarded a passing grade or a failing grade.

Table 6.3.2 presents these two rater agreement figures for the writing section in 2009 and 2010.

Table 6.3.2 Writing Section Rater 1 / Rater 2 and Pass / Fail Agreement for the 2009 and 2010 Administrations

	Rater 1 / Rater 2 Agreement (%)	Pass/Fail Agreement (%)
MJ2009	74.0	87.5
N2009	77.7	90.7
MJ2010	76.7	84.7
N2010	78.8	88.5

The trend is generally upwards for R1/R2 agreement, and though the pass/fail agreement is a little less predictable, it is at a good level. CaMLA is working with the examiners for the writing section to continue the upwards trend.

The examiners for the ECPE speaking test are native or highly proficient nonnative speakers of English, all trained and certified according to standards set by CaMLA. There are two examiners present during the speaking test and they separately award a score to each test taker. At the end of the speaking test, the examiners discuss their individual ratings and arrive at a consensus on the final score to be awarded. It is important to note that *the final score awarded for each test taker is the result of a discussion by both speaking test examiners who have each independently evaluated the speaking performance and then arrive at a consensus rating*. This means that no single rater can determine the final outcome for a test taker.

The agreement between examiners is carefully monitored. CaMLA monitors exact agreement between the first and second rater. Table 6.3.3 presents the rater agreement figures for the Speaking test in 2009 and 2010.

Table 6.3.3 Speaking Test Rater 1 / Rater 2 Agreement for the 2009 and 2010 Administrations

Administration	Rater 1 / Rater 2 Agreement (%)
MJ2009	90.12
N2009	89.40
MJ2010	88.90
N2010	88.80

The agreement figures are above 0.80 and are acceptable. However, the *rater 1 / rater 2* agreement figures indicate a slight downward trend. This suggests

examiner drift and needs to be addressed quickly to ensure that the rater agreement figures stabilize. CaMLA is working with examiner trainers to help speaking test examiners recalibrate.

6.4. Standard Errors

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

Table 6.4.1 SEM Estimates for the Listening and GCVR Sections During 2009 and 2010

Section	MJ2009	N2009	MJ2010	N2010
Listening	0.43	0.40	0.38	0.39
GCVR	0.27	0.28	0.27	0.27

Table 6.4.1 presents the SEM estimates in logits for the listening and GCVR sections during 2009 and 2010. It shows that the SEM estimates are generally stable. The small drop in the SEM figures for the listening sections in the 2010 administrations suggests a slight increase in the precision of measurement for those administrations.

6.5. 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 ECPE captures evidence of a test taker's communicative language ability at the C2 level of the CEFR through a variety of tasks in the four language skills of listening, reading, speaking, and writing. Section 3.4 describes the skills and abilities expected for each language skill as well as for grammar and vocabulary knowledge. Even though performance on the ECPE is expressed as a pass or fail—that is, a test taker has to pass the ECPE 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 (see the appendix). Reporting scores in this way is justifiable if each section can be seen to contribute differentially to the overall ECPE result. Table 6.5.1 presents the subtest correlations (Spearman's rho) for the 2009 and 2010 administrations.

Table 6.5.1 Subtest Correlations (Spearman rho) for the 2009 and 2010 Administrations[4]

		Speaking	Writing	Listening
MJ2009	Writing	0.284*		
	Listening	0.377*	0.306*	
	GCVR	0.454*	0.432*	0.626*
N2009	Writing	0.307*		
	Listening	0.378*	0.290*	
	GCVR	0.435*	0.416*	0.600*
MJ2010	Writing	0.292*		
	Listening	0.377*	0.342*	
	GCVR	0.443*	0.451*	0.649*
N2010	Writing	0.278*		
	Listening	0.368*	0.328*	
	GCVR	0.435*	0.422*	0.649*

The correlations range between 0.3 and 0.6, indicating a moderate relationship between the subtests. This is to be expected since each subtest intends to capture language proficiency from a different perspective.

4 * Correlation is significant at the 0.01 level (2-tailed).

7. Additional ECPE Validity Evidence

Section 3.4 (above) presented a proposed interpretation of the ECPE certificate. 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 ECPE along with the research evidence available for these claims.

Table 7.1 Proposed Validity Claims About the ECPE and the Research Evidence Available

Proposed claim	Research evidence available
The different item types and tasks are appropriate for measuring language proficiency at the C2 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, Ann Arbor, MI: University of Michigan.
The structure of the ECPE is consistent with its stated construct and with the way in which scores are reported.	• Ameriks, Y. (2009) *Investigating Validity Across Two Test Forms of the Examination for the Certificate of Proficiency in English (ECPE): A Multi-Group Structural Equation Modeling Approach*. Dissertation Abstracts International Section A: Humanities and Social Sciences, 70:21–A. • Ameriks, Y. (2010) *Investigating the dimensionality of grammatical knowledge and reading ability across two test forms*, paper presented at the Language Testing Research Colloquium, Cambridge, UK, April 2010. • Römhild, A. (2008). *Investigating the Invariance of the ECPE Factor Structure across Different Proficiency Levels*, Spaan Fellow Working Papers in Second or Foreign Language Assessment. • Wang, S. (2006). *Validation and Invariance of Factor Structure of the ECPE and MELAB across Gender*, Spaan Fellow Working Papers in Second or Foreign Language Assessment.
The language elicited by the speaking and writing sections of the ECPE reflects the language expected at the C2 level on the CEFR.	• Plough, I.C., MacMillan, F., and O'Connell, S.P. (2011) Changing Tasks … Changing Evidence: A Comparative Study of Two Speaking Proficiency Tests, in Granena, G., Koeth, J., Lee-Ellis, S., Lukyanchenko, A., Botana, G.P., and Rhoades, E. (Eds), *Selected Proceedings of the 2010 Second Language Research Forum*, Somerville, MA: Cascadilla Proceedings Project. 91-104.
The ECPE has positive consequences for stakeholders.	• Wang, S. (2006). *Validation and Invariance of Factor Structure of the ECPE and MELAB across Gender*, Spaan Fellow Working Papers in Second or Foreign Language Assessment. • Zhang, B. (2010) Assessing the accuracy and consistency of language proficiency classification under competing measurement models, *Language Testing*, 27(1): 119–140.

7.1. The different item types and tasks are appropriate for measuring language proficiency at the C2 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 (ECCE) and the ECPE. 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 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 ECPE listening and reading section items and tasks appropriately target the level of the test. They have informed revisions to item specifications, ongoing item development, and test construction. In order to confirm the success of these efforts, 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 items are measuring at the C2 level on the CEFR. A standard-setting study that explicitly investigates this question has been prioritized.

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

Wang (2006) used factor analysis to validate the internal structure of the ECPE and the Michigan English Language Assessment Battery (MELAB) The ECPE data comprised 2011 test takers from one administration of the speaking, listening and GCVR (reading) sections. A series of analyses were conducted beginning with descriptive statistics, internal consistency, and intercorrelations of subtests and tests. These initial analyses revealed that there were unequal N-counts for male and female test takers; about 59% of the test takers in this dataset were female, a finding that is in line with the trends described in section 5.3. The analyses also showed that the male test takers had

slightly higher mean test score than the female test takers, in line with the findings reported in section 6.2.

Second, Wang performed exploratory factor analyses (EFA) that took into account the different item types represented in the exam. These analyses revealed one dominant factor with an eigenvalue of 2.20, which accounted for 91.7% of the common variance. This result suggested that the EPCE data was unidimensional with one underlying construct–language proficiency. Wang confirmed this by comparing more closely the differences in the eigenvalues between the first and second factors with the differences in the eigenvalues between the second and third factors. According to Hattie (1985, cited in Wang, 2006: 45), if a test is unidimensional, then the ratio of these differences will be large. The ratio for the ECPE was 37.7, confirming a single meaningful factor to explain the ECPE data.

After determining that a one-factor model best explained the data, Wang performed a confirmatory factor analysis (CFA) that was cross-validated by splitting the dataset into two randomly assigned samples—a calibration sample and a validation sample. Wang found that most of the fit statistics were acceptable and concluded that the total score for the ECPE (when all the sections are taken together in the calculation of a test taker's exam result) measures English language proficiency. This supports claims that the ECPE test sections together measure a test taker's overall English language proficiency as well as CaMLA's practice of awarding an ECPE certificate based on the test takers' performance on all sections of the examination.

In another investigation of the way in which scores are reported, Römhild (2008) investigated the factor structure of the ECPE across different proficiency levels. She argued:

> If the factor structure of a test varies across examinees as a function of their language proficiency level, then score comparisons are no longer meaningful because different kinds of information are gained from the test score for different groups of examinees. As a consequence, test developers may need to rethink the use of composite scores for multiple language skills and knowledge components.
>
> Römhild (2008: 30)

Römhild's (2008) dataset comprised the listening and reading sections of one administration of the ECPE (N = 34,599). Her first task was to identify the proficiency groups. In the absence of independent information to divide the sample into low- and high-proficient groups, Römhild split the test items into two test halves of odd- and even-numbered items and determined that the two halves yielded approximately equivalent score distributions. She also determined that the mean difference (in scores) between the two proficiency groups identified was statistically significant.

Römhild then conducted exploratory factor analyses to identify the latent factors that best accounted for the data in one test half (the even-numbered items—ECPE (even). Like Wang (2006), she found a strong first factor and multiple secondary factors (2008: 34). She found that the ratio of the difference in the eigenvalues between the first and second factors with the difference in the eigenvalues between the second and third factors was 9.96, confirming a single meaningful factor to explain the data. Diverging from Wang's (2006) findings, however, Römhild also identified a promising three-factor model comprising listening and vocabulary as separate factors and a third factor that combined grammar, cloze, and reading items. To these two empirically derived models she added a third, five-factor model that represented the subtest structure of the ECPE; that is, listening, grammar, cloze, vocabulary, and reading.

Römhild compared all three models for goodness of fit for both proficiency groups (low- and high-proficiency test takers). Though none of the models met all the fit criteria, she found that the five-factor model provided the best model fit for both proficiency groups and also exhibited adequate model fit as measured by the RMSEA (Root Mean Square Error of Approximation) and the χ^2/df ratio. Römhild then checked whether the same construct is measured in the low- and high-proficiency groups and found that the factor structure of the ECPE differed between the two groups and that, as proficiency increases, ability in individual skills (e.g., listening, grammar, vocabulary, etc.) converges into general language proficiency. As a result of her analyses, Römhild warns that the composite scores for the low- and high-proficiency groups carry different meanings.

Römhild (2008) arrived at a very different conclusion from Wang (2006). While Wang's work supports CaMLA's practice of awarding an ECPE certificate based on the test takers' performance on all sections of the examination, Römhild's findings strike a note of caution for test users. However, she identifies a number of limitations of her study. First, she admits

that she failed to cross-validate her findings with the odd-numbered items. Second, she proceeded with a five-factor model that displayed adequate fit on two measures but not on all measures. In other words, she is not able to be confident of the appropriateness of the five-factor model. Thirdly, this five-factor model does not reflect the way in which ECPE scores are reported and therefore does not represent CaMLA's claims about the ECPE construct. Finally, she divided the dataset into only two proficiency groups. This could have resulted in muddying of the analyses at the boundary between the two groups.

Ameriks (2009 and 2010) applied structural equation modeling to the reading section of two ECPE forms. Her aim was to first determine the underlying constructs of the section and then to examine the extent to which these constructs were invariant across different test forms. She argued that it is insufficient to compare test scores across forms or to apply equating procedures to ensure that the scores reported are population independent. It is also important to supply empirical evidence of equivalence of construct.

Ameriks' dataset comprised the performance data from the 2003–2004 and 2004–2005 administrations of the ECPE reading section (N = 66,135). The grammar, cloze, and vocabulary items were coded according to Purpura's (2004) model of grammatical knowledge. The reading items were coded separately for the reading subskill they tested. The variables identified through this analysis were used to hypothesize a construct model for the reading section, which was then tested both within each test form and across the test forms. Ameriks cautions her readers that the theoretical models of lexico-grammatical knowledge and reading proposed in her study may not be the only models (or even the best models) to explain the data. She also warns that the labels assigned to the factors might be subject to the *naming fallacy* (Kline, 1998). Nevertheless, her investigations indicate that, within the models that she has posited, both test forms conform to a two-factor model of grammatical knowledge and reading ability with nine measured variables. Though her tests of invariance across the forms revealed differences that were statistically significant, the actual values were very similar. Ameriks suggested a number of explanations for the differences in the parameter estimates, such as differences in the population that took the two forms, errors in the coding, and differences in the particular grammatical features tested in each form. She suggested, therefore, that "the parameter estimates were substantively equivalent" (2009: 183).

Ameriks' (2009) work shows that the structure of the ECPE reading section is consistent with its stated construct; that is, to test lexico-grammatical knowledge and reading. Her research also shows that the construct of the ECPE reading section is stable across different forms of the exam. Lessons learned from Ameriks' coding have been incorporated into more recent iterations of the ECPE form specifications, and revisions have been made to the cloze subsection. It would therefore be worth repeating her analyses with more recent test forms. It would also be useful to apply her methodology to the listening section.

7.3. The language elicited by the speaking and writing sections of the ECPE reflects the language expected at the C2 level on the CEFR

As part of the introduction of the new ECPE speaking test in May 2009, Plough et al. (2011) examined the language elicited by the new test in order to confirm the construct claims made about the test. The data comprised performances by 39 ECPE test takers on both forms of the speaking test (the old format—which was last administered in November 2008—and the new format). The performances were collected within two weeks of one another, ensuring little language gain in the intervening period. The performances were transcribed and then analyzed for language complexity (as measured by AS-units), vocabulary range, and range of language functions. The results of this analysis were heartening. They showed that both the old and the new versions of the speaking test elicited similar levels of language complexity and a comparable range of vocabulary. This confirmed that the old speaking test, though narrower in its construct, gave ECPE test takers opportunities to demonstrate their language proficiency. This is important in terms of the continued meaningfulness of older ECPE scores. However, the old and new versions of the speaking test elicited markedly different language functions. The new ECPE speaking test tends to elicit a wider variety of language functions and also more complex, embedded functions. This, in turn, suggests that the new ECPE speaking test has a broader construct than its predecessor.

As it stands, Plough et al.'s (2011) work provides some evidence of the language elicited by the ECPE

speaking test but does not explicitly address the question of whether the elicited language reflects that expected of a test taker at the C2 level on the CEFR. It would be useful to extend their work by comparing it to the claims made for C2 speaking proficiency on the CEFR (cf., Council of Europe, 2001: 28-29). Additionally, since the data is held at CaMLA, its use should be maximized by performing more detailed analyses, both in terms of the measures applied (e.g., measures of vocabulary richness, formulaic sequences, more sophisticated morpho-syntactic analyses) and in terms of how the data is dissected or transected. For instance, Plough et al. (2011) looked at each performance as a unit. Further analyses could investigate each stage of the test separately to accumulate further validity evidence for the five-stage design. Performances could also be analyzed cross-sectionally to validate the rating scale.

There are no published investigations of the language elicited by the writing section of the ECPE. This would be welcomed: particularly studies that explore the equivalence of writing prompts across different ECPE administrations and also studies that analyze the language elicited by the ECPE writing section.

7.4. The ECPE has positive consequences for stakeholders

Investigations into the consequences of the ECPE have been varied. Wang (2006) examined the factor structure of the ECPE across gender. Wang's (2006) dataset comprised 2011 test takers (59% female) from one administration of the speaking, listening and GCVR (reading) sections. Even though the male test-taker group had a higher mean score for the exam than the female test-taker group, the factor analyses demonstrated that the "models for male and female students [had] structure, factor loading, and variance equivalence" (Wang, 2006: 53). This, in turn, suggests that the test is fair across gender groups.

Zhang (2010) investigated the accuracy of the pass/fail classification decision using four different measurement models: classical test theory (CTT); dichotomous item response theory (IRT); testlet response theory (TRT); and, polytomous item response theory (Poly-IRT). Zhang's dataset comprised 5,000 randomly selected test takers from one administration of the listening and reading sections. He first inspected the items, excluding three (two listening items and one vocabulary item) because they had negative corrected point-biserial correlations. He then calculated the Cronbach's alpha coefficients of each section in order to ensure that they met minimum requirements. Next, Zhang applied each of the measurement models to the data. When he looked at the reading and cloze subtests within the reading section, Zhang found differences in classification accuracy which suggested a strong testlet effect. This is unsurprising given that these subtests comprise sets of items that are related to a single prompt.

It should be noted that the test takers' performance on subtests are not reported. Instead, the results of the main ECPE sections are reported (in this case listening and reading). Therefore, Zhang's analysis of the classification decision accuracy for the listening and reading sections was more interesting and relevant to the interpretation of ECPE scores. Here Zhang found that the classification decisions were similar for the different measurement models and that the overall classification accuracy was high for both sections. Zhang's view was that the TRT model was the superior of the four models applied—primarily because he found strong testlet effects within the cloze and reading subtests. He argued that the IRT model overestimates the accuracy of the proficiency classification. Nevertheless, his results show that the classification accuracy for the TRT model (84.8) and the IRT model (85.0) were virtually identical and the classification consistency between these models was high (98.4). Additionally, only 30% of the items in the listening section and only 33% of the items in the reading section are in testlets. These two points taken together suggest that the IRT model is sufficiently accurate for the ECPE. Additionally, as Zhang has conceded, the TRT procedure is much harder and more time consuming to apply—an important consideration for large-scale testing programs where exam results must be issued in a timely manner.

The research reported here shows that there are a number of ways in which to ensure that the ECPE has positive consequences for stakeholders. Wang (2006) used factor analysis to show that exam is fair across gender groups. Zhang's (2010) analyses showed that the measurement model adopted for the ECPE is appropriate. There is also work in progress (Tsagari, forthcoming) that studies the influence of the ECPE upon ECPE-focused textbooks. Additional research showing the stability of the factor structure of the exam over time and stability of item performance over time

would be useful. It would also be useful to conduct follow-up studies with test takers to see how they have benefited from studying for the exam.

The research already completed has made substantial progress towards building a validity argument for the ECPE. However, proposals would be welcomed for further research, particularly work that could support the following claims about the ECPE:

- The content of the test is representative of the kinds of oral and written texts and tasks that might be encountered by very advanced learners of English at the C2 level on the CEFR.
- The writing and speaking rating scales reflect the features of language proficiency expected of learners of English at the C2 level on the CEFR.
- The language processes and linguistic knowledge that the test takers use to successfully complete the ECPE reflects the language knowledge and processes expected at the C2 level on the CEFR.
- ECPE test results are used appropriately.

8. References

Ameriks, Y. (2009) *Investigating Validity Across Two Test Forms of the Examination for the Certificate of Proficiency in English (ECPE): A Multi-Group Structural Equation Modeling Approach*. Dissertation Abstracts International Section A: Humanities and Social Sciences, 70:21–A.

Ameriks, Y. (2010) *Investigating the dimensionality of grammatical knowledge and reading ability across two test forms*, paper presented at the Language Testing Research Colloquium, Cambridge, UK, April 2010.

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.

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 Publications Inc.

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

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

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

Kline, R. B. (1998) *Principles and practices of structural equation modeling*, New York, NY: The Guilford Press.

Leckie, G. and 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.

Plough, I. C., MacMillan, F., and O'Connell, S. P. (2011) Changing Tasks . . . Changing Evidence: A Comparative Study of Two Speaking Proficiency Tests, in Granena, G., Koeth, J., Lee-Ellis, S., Lukyanchenko, A., Botana, G.P., and Rhoades, E. (Eds), *Selected Proceedings of the 2010 Second Language Research Forum*, Somerville, MA: Cascadilla Proceedings Project. 91–104.

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

Römhild, A. (2008). *Investigating the Invariance of the ECPE Factor Structure across Different Proficiency Levels*, Spaan Fellow Working Papers in Second or Foreign Language Assessment.

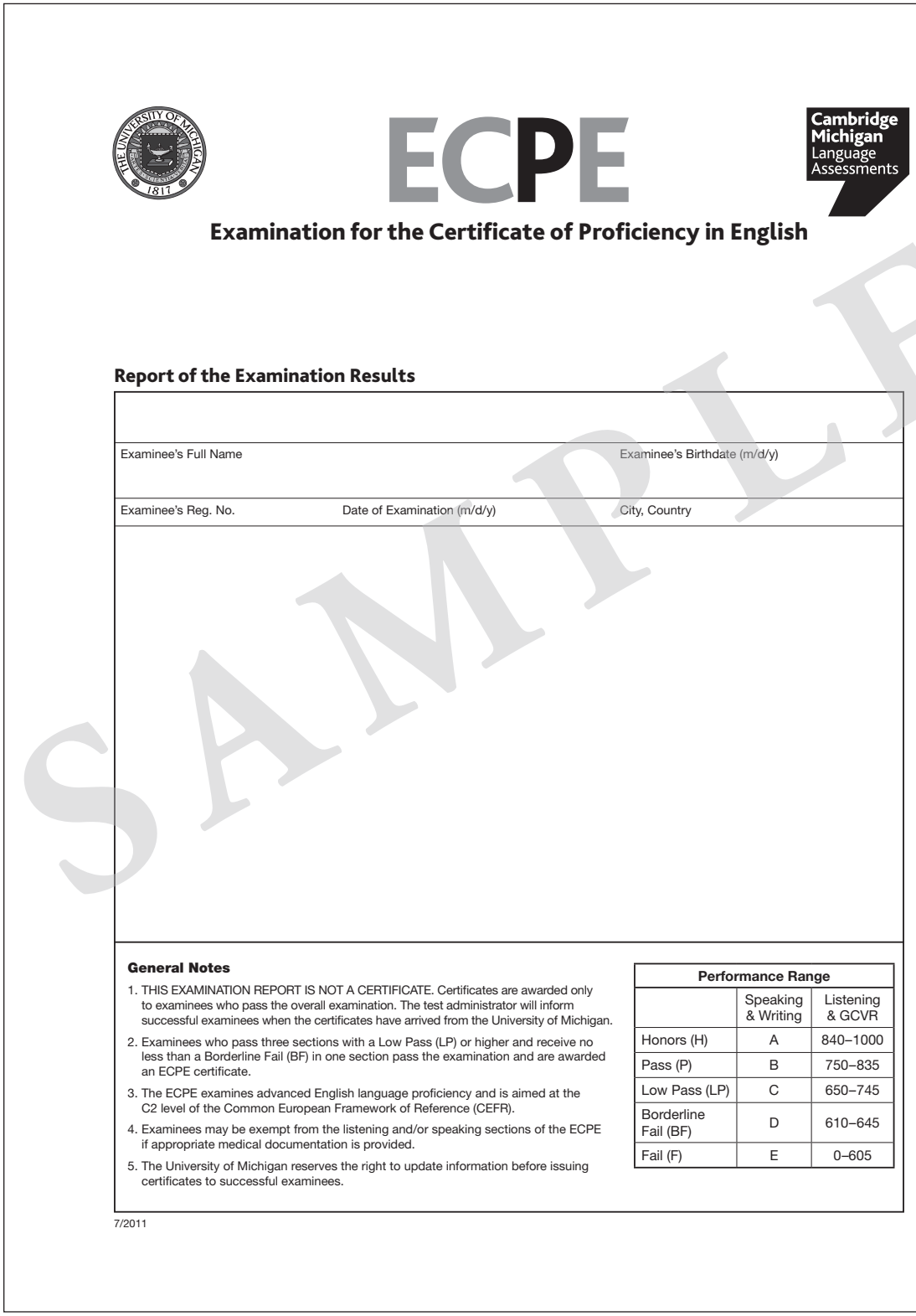
Tsagari, D. (forthcoming) *Writing to the test: the influence of the ECPE upon test preparation textbooks*, Spaan Research Grant, awarded 2010.

Wang, S. (2006). *Validation and Invariance of Factor Structure of the ECPE and MELAB across Gender*, Spaan Fellow Working Papers in Second or Foreign Language Assessment.

Zhang, B. (2010) Assessing the accuracy and consistency of language proficiency classification under competing measurement models, *Language Testing*, 27(1): 119–140.

Appendix

Sample Examination Report



ECPE

Cambridge Michigan Language Assessments

Examination for the Certificate of Proficiency in English

Report of the Examination Results

Examinee's Full Name		Examinee's Birthdate (m/d/y)
Examinee's Reg. No.	Date of Examination (m/d/y)	City, Country

General Notes

1. THIS EXAMINATION REPORT IS NOT A CERTIFICATE. Certificates are awarded only to examinees who pass the overall examination. The test administrator will inform successful examinees when the certificates have arrived from the University of Michigan.
2. Examinees who pass three sections with a Low Pass (LP) or higher and receive no less than a Borderline Fail (BF) in one section pass the examination and are awarded an ECPE certificate.
3. The ECPE examines advanced English language proficiency and is aimed at the C2 level of the Common European Framework of Reference (CEFR).
4. Examinees may be exempt from the listening and/or speaking sections of the ECPE if appropriate medical documentation is provided.
5. The University of Michigan reserves the right to update information before issuing certificates to successful examinees.

Performance Range		
	Speaking & Writing	Listening & GCVR
Honors (H)	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

7/2011