2009–2013
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
www.CambridgeMichigan.org

6/2014

Table of Contents

List of Tables

1. Introduction

The Michigan English Test (MET) is a test of general language proficiency for learners of English. It is administered monthly at test centers around the world. From 2009 to 2013, the exam was administered 56 times, a minimum of 11 times each year.

This report provides test users with technical information about the MET. Section 2 provides general information about the test and a proposed interpretation of MET 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 MET from 2009 to 2013. The remaining parts of the report focus on statistical analyses of test data from the five year period. Section 5 discusses the MET test taking population, looking particularly at the yearly distributions of the test takers by gender and age. Section 6 looks at trends in the MET 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 MET results.

2. Description of the MET

2.1. General Description

The MET is a standardized multilevel examination of general English language proficiency. It measures listening, reading, grammar, and vocabulary skills in personal, public, occupational, and educational contexts. Listening recordings and reading passages reflect authentic, everyday interaction in an English speaking environment. As of 2013, an MET Speaking Test is also available.

The MET covers a range of proficiency levels from upper beginner to lower advanced; the A2 to C1 levels of Common European Framework of Reference (CEFR), with emphasis on the middle range of B1 and B2. The MET is intended for adults and adolescents at or above a secondary level of education who want to measure their general English language proficiency in a variety of linguistic contexts. The MET can be used for educational purposes, such as when finishing an English language course, or for employment purposes, such as applying for a job or pursuing promotion that requires an English language qualification.

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 MET is open to all people who wish to take the exam, regardless of the school they attend.

2.2. Proposed Interpretation of the Scores

The MET is a multilevel exam, covering proficiency levels A2 to C1 on the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (Council of Europe, 2001) (CEFR). If test takers are at the A2 level they are considered to be basic users of English. They are generally able to understand short conversations, particularly if the speakers talk slowly and clearly. They can communicate through speaking on routine tasks that require a direct exchange of information. They can also read short, simple texts about topics that are familiar to them. Their vocabulary typically covers topics related to personal and family information, shopping, and other matters of personal relevance.

Test takers at the B1 and B2 levels are independent users of English. Test takers at the B1 level can understand the main points of conversations and short talks, as long as the topic is familiar and the presentation is well-structured. They can communicate through speaking on a variety of topics related to their hobbies and their job. They can also read well-structured, factual texts. Their vocabulary typically covers topics related to their personal life and their work. Test takers at the B2 level have a broader and deeper grasp of English than those at the B1 level. They can 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.

Test takers at the C1 level are considered to be proficient users of English. They can follow conversations with ease and are also able to understand most lectures, discussions, and debates, even when the interaction is lively. They speak fluently and appear to make no effort to formulate their thoughts. They can also read lengthy and complex texts on a variety of topics. They are able to make inferences about the writer's opinion and attitudes as well as understand fine details. Their vocabulary is very broad.

2.3. Test Structure

The MET measures listening, reading, grammar, vocabulary, and speaking skills. Table 2.3.1 describes the format and content of the MET. Test preparation resources are available on the CaMLA website.

3. Scoring and Reporting of MET Results

3.1. Explanation of Scoring for Each Section

The speaking test is conducted and assessed by one examiner who has been trained and certified according to CaMLA standards. The speaking performances are graded according to a scale established by CaMLA (see the CaMLA website for the speaking rating scale). The listening and reading sections are scored by computer at CaMLA. Each correct answer contributes proportionally to the final score and there are no points deducted for wrong answers. A scaled score, ranging from 0 to 80, is calculated using an advanced mathematical model based on Item Response Theory

Table 2.3.1 Format and Content of the MET

Section	Description	Number Of Items	Time
Listening	Multiple-choice questions that assess the ability to understand conversations and talks in three parts. **Part 1** Short conversations **Part 2** Longer conversations **Part 3** Talks	60 total	45 minutes
Reading and Grammar	**Two Parts** Multiple-choice questions testing a variety of grammar structures. Multiple-choice questions that assess the ability to understand a variety of passages in social, educational, and workplace contexts. *Vocabulary is assessed within the listening and reading sections.*	75 total • 25 grammar • 50 reading	90 minutes
Speaking	A structured one-on-one interaction between an examiner and a test taker.	5 tasks	10 minutes

(IRT). This method ensures that the ability required to achieve a particular scaled score remains the same from year to year and that scores are comparable across different administrations.

3.2. Equating Procedures

In order to ensure that MET 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 MET to be equated. The scale and location parameters are computed separately for each section of the exam.

3.3. Procedures for Reporting Scores

All test takers receive a CaMLA score report that shows their scaled score for each section, ranging from 0 to 80. The section scores are also reported as a CEFR level: A2-C1. Table 3.3.1 shows the MET scaled scores that correspond to each CEFR level. These correspondences are based on standard setting research conducted by CaMLA. A report documenting the process can be found on the CaMLA website.

Table 3.3.1 CEFR Level Equivalence of the MET Scaled Scores

CEFR Level	Scaled Score
C1	64–80
B2	53–63
B1	40–52
A2	0–39

The score report includes a final score, which is the total of the listening and reading sections of the test. The speaking score is not included in the final score; rather, it is reported separately on the score report. There are no CEFR correspondences provided for the final score.

3.4. Interpretation of Scores for Each Section

The score reports provide a brief description of what a test taker can do at each level. This allows score users to easily see what test takers at a particular CEFR level can do in English. When interpreting MET results, it is important to remember that the MET estimates a test taker's true proficiency by approximating the kinds of tasks that may be encountered in real life. Also, temporary factors unrelated to test takers' proficiency, such as fatigue, anxiety, or illness, may affect exam results.

3.5. Guidelines for Decision-Making

When using test scores for decision making, look at each section score separately. It is possible for a test taker to be at a higher language proficiency level in one language skill than in another. Therefore, all section scores should be taken into account when interpreting the test results for use in decision making. Additionally, check the date the test was taken. While the MET report is valid for two years, language ability changes over time. This ability can improve with active use and further study of the language, or it may diminish if the report 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 knowledge and skills such as intellectual knowledge and business skills.

4. Changes to the MET from 2009–2013

In the period covered by this report, three changes have been introduced for the MET. One was a change to the design of the test. The other two were scoring changes.

4.1. Changes to the Design of the Test

Addition of Speaking Test

Until January 2013, the MET consisted of only two sections, one testing listening and one testing reading and grammar. A speaking section was added in 2013.

The speaking test measures an individual's ability to produce comprehensible speech in response to a range of tasks and topics. It is a structured one-on-one interaction between examiner and test taker that includes five distinct tasks. The tasks require test takers to convey information about a picture and about themselves, give a supported opinion, and state the advantages and disadvantages of a particular proposal. More details on the speaking section can be found on our website.

Information about how the MET speaking test performed during its first year of administration is available in the MET 2013 Report (http://www.cambridgemichigan.org/sites/default/files/resources/Reports/MET-2013-Report.pdf).

4.2. Scoring Changes

Changes in Reading Section Cut Score

The cut scores for the listening and reading sections were established by an empirical standard setting study using judgments from experienced English language teachers and administrators. During the first year of test administrations, the cut scores were monitored to ensure that they were appropriate. The initial MET A2/B1 cut score for the listening section was a scaled score of 40 while the A2/B1 cut score for the reading section was a scaled score of 37. The monitoring revealed that the A2/B1 cut score for listening was appropriate but the reading one was not. As of January 2010, the A2/B1 cut score for the reading was changed to 40; this aligned it with the cut score for the listening section.

Changes in Range of Scaled Scores

Until 2011, the MET reported scaled scores on the range 0 to 100, with the cut scores at 40, 53, and 64 for the A2/B1, B1/B2, and B2/C1 CEFR levels, respectively. During the first two years of test administration, test taker performances were monitored for use of this score range. The data showed that a perfect raw score typically corresponded to a scaled score at or below 80. This had the potential to cause a misinterpretation for score users. Test takers that performed well may wonder why they did not receive a score closer to the maximum score of 100. Additionally stakeholders might set expectations for test takers to achieve scores above 80 when, in reality, this could not occur. As a result, starting in January 2011, the range of scaled scores was truncated to the current 0 to 80 scale with the same CEFR cut scores.

5. MET Test-Taking Population

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

5.1. Test-Taker Numbers

Table 5.1.1 presents the number of test takers who took the MET each year, from 2009 to 2013. It shows that the number of test takers has increased dramatically in the past 5 years. From 2012 to 2013 alone, the number of test takers increased by over 3,000, a 35% growth in the population.

Table 5.1.1 Number of MET Test Takers

Year	Number
2009	5,192
2010	6,767
2011	7,050
2012	8,832
2013	11,975

5.2 Administering Countries

In 2009, when the MET was launched, it was administered in only one country, Colombia, but by 2013 the MET program had expanded to twelve countries. Table 5.2.1 lists the different countries that have administered the MET. The test centers are located in several different geographic areas; CaMLA would like to increase the number of test centers in all parts of the world.

Table 5.2.1 Administering Countries

Albania	Colombia	Mexico
Argentina	Costa Rica	Peru
Brazil	Croatia	Serbia
Chile	Italy	United States

5.3. Language Distribution

Every MET 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. The 2013 MET was taken by test takers of 16 different first language backgrounds. Table 5.3.1 presents the distribution (in %) of test taker first languages throughout the five year period. Although the majority of the test takers list Spanish as their first language, the table shows that the percentage of test takers speaking a different language is increasing. This growth can be attributed to increases in the number of countries that administer the MET.

Table 5.3.1 Distribution (in %) of Test-Taker First Language*

	2009	2010	2011	2012	2013
Albanian	0.00	0.00	0.00	7.01	16.73
Portuguese	0.05	1.00	1.53	0.51	2.19
Spanish	99.33	98.75	98.06	92.11	80.36
Others	0.62	0.25	0.41	0.37	0.72

* Percentages are reported for language groups with more than 50 test takers in 2013.

5.4 Gender Distribution

The MET 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 5.4.1 shows the distribution of test takers by gender from 2009 to 2013. It shows that more females take the test than males each year. Over time, this difference has become more pronounced.

Table 5.4.1 Distribution (in %) of MET Test Takers by Gender

Year	Male	Female	Missing Data
2009	44.60	48.70	6.70
2010	45.81	52.96	1.23
2011	45.10	54.90	0.00
2012	42.74	56.43	0.83
2013	41.88	57.84	0.28

5.5 Age Distribution

The MET 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 5.5.1 shows the distribution of test takers by age for each year. The table shows that, each year, the majority of MET test takers are between 17 and 25 years old. This suggests that the majority of MET test takers registered for the exam while still attending a school or university, or in the early stages of their careers.

Table 5.5.1 Distribution (in %) of MET Test Takers by Age

Age	2009	2010	2011	2012	2013
≤ 12	0.00	0.31	0.41	0.60	0.48
13–16	7.78	10.57	15.35	15.55	10.64
17–19	24.90	22.24	17.66	15.60	15.22
20–22	14.41	15.10	17.36	20.01	22.41
23–25	18.89	20.30	18.61	20.46	22.71
26–29	13.14	13.77	13.30	11.85	11.92
30–39	11.63	12.78	13.46	11.54	10.33
≥ 40	5.74	4.54	3.84	3.71	5.44
Missing Data	3.51	0.38	0.00	0.68	0.84

Table 5.5.1 also shows a slight upward trend in the number of test takers in the older age groups. We investigated this further by looking at the average age of the population. Table 5.5.2 shows the average age of test takers for the three major language groups, as well as for the overall test taking population. From

this table, we can see that the average age of the test taking population has increased slightly in 2013. We can also see that the Albanian speaking test takers are, on average, several years older than the remainder of the population. This difference, combined with the increased number of Albanian speaking test takers (see Table 5.3.1, above) has probably led to the increase in the average age of the test taking population that was identified in Table 5.5.1.

Table 5.5.2 Average Age of MET Test Takers by Language

Language	2009	2010	2011	2012	2013
Albanian	NA	NA	NA	27.24	28.94
Portuguese	21.00	13.05	17.32	14.50	20.81
Spanish	23.20	23.42	23.73	23.13	23.07
Overall	23.55	23.38	23.64	23.34	24.03

6. MET Results and Test Statistics

This section presents the results of the test takers who took the MET during the period covered by this report. It examines trends in the distribution of scores by age and gender, as well as trends in reliability and SEM.

6.1. Trends in CEFR Level Distribution

Listening

Table 6.1.1 shows the distribution (in %) of MET test-taker CEFR levels from 2009 to 2013 for the listening section of the exam. The table shows that the majority of the test takers score at the B1 level each year. The second largest percentage of test takers is at the B2 level prior to 2011, and then at the A2 level after 2011. This change in listening proficiency distribution corresponds with the growth in test taker numbers and increase in the number of countries in which the MET is administered.

Table 6.1.1 Distribution (in %) of Each CEFR Level on the Listening Section

Year	A2	B1	B2	C1
2009	19.75	41.49	26.42	12.33
2010	21.41	42.69	24.96	10.94
2011	22.43	43.03	23.41	11.13
2012	23.74	43.34	23.38	9.53
2013	27.12	42.41	21.01	9.46

Reading

Table 6.1.2 shows the distribution (in %) of MET test taker CEFR levels from 2009 to 2013 for the reading section of the exam. This section reveals a trend similar to the listening section. Over this five year span, the majority of the test takers scored at the B1 level of the exam. The second largest percentage of test takers is at the B2 level prior to 2013, and then at the A2 level in 2013. Like the listening section, as the MET test population has grown and diversified in terms of the first language of the test takers, it has also changed in its reading proficiency distribution.

Table 6.1.2 Distribution (in %) of Each CEFR Level on the Reading and Grammar Section

Year	A2	B1	B2	C1
2009	0.33	32.61	44.58	22.49
2010	10.94	49.16	31.85	8.05
2011	15.65	50.74	27.17	6.44
2012	22.21	47.04	24.72	6.02
2013	25.22	46.23	21.90	6.61

Overall, the data show that the test taking population is trending towards lower CEFR levels. The most likely explanation of this trend is related to the increased number of test takers and administering countries. As previously discussed in Sections 5.1 and 5.2, in 2009 the MET was administered to 5,192 test takers in one country, whereas in 2013, it was administered to 11,975 test takers in twelve countries.

The substantial increase in the number of test takers and the number of administering countries has probably added a wider variety of educational and cultural backgrounds and skill levels. To explore this hypothesis, we examined the distribution of the CEFR

levels for each of the three primary first language groups in 2012 and 2013. This corresponds to the entry of the Albanian test taking population (who represent the second-largest language group in the test population). Table 6.1.3 presents the distribution (in %) of the test takers by L1 and CEFR level for the listening section. Table 6.1.4 presents the distribution (in %) of the test takers by L1 and CEFR level for the reading and grammar section.

Table 6.1.3 Distribution (in %) of Each CEFR Level by First Language on the Listening Section

Year		A2	B1	B2	C1
2012	Albanian	38.31	52.55	8.26	0.88
	Portuguese	9.76	53.66	26.83	9.76
	Spanish	22.10	42.35	25.04	10.50
2013	Albanian	46.37	39.62	11.85	2.16
	Portuguese	17.32	38.98	34.65	9.06
	Spanish	24.32	43.10	22.09	10.50

Table 6.1.4 Distribution (in %) of Each CEFR Level by First Language on the Reading and Grammar Section

Year		A2	B1	B2	C1
2012	Albanian	47.80	45.17	6.33	0.70
	Portuguese	7.32	68.29	19.51	4.88
	Spanish	21.65	46.06	25.91	6.38
2013	Albanian	48.27	39.52	10.66	1.55
	Portuguese	15.35	42.13	30.31	12.20
	Spanish	21.46	47.96	23.55	7.03

The tables show that the Portuguese-speaking test takers are similar in profile to the Spanish-speaking test takers. However, the Albanian-speaking test takers typically take the MET when they are at lower CEFR levels compared to the other major language groups. This suggests that the changes in the CEFR distributions observed in Table 6.1.2 are the results of changes in the nature of the test taking population.

6.2. Trends in CEFR Level Distribution by Age and Gender

Table 6.2.1 presents the overall distribution (in %) of the CEFR levels for each gender in the listening and reading sections of the MET. The data seems to suggest that, for both sections, a larger percentage of the males obtain C1 and B2 levels, while a larger percentage of the female population obtain B1 and A1 levels. This implies that the males tend to score higher than females on this exam. In order to determine whether these differences were meaningful, we ran a cross-tabulation and chi-squared test for the entire test population.

Table 6.2.1 Overall Percentage of Male and Female Test Takers Receiving Each CEFR Level for Both Sections

Section	Gender	A2	B1	B2	C1
Listening	Male	19.96	42.20	26.02	11.82
	Female	26.35	43.42	21.21	9.03
Reading	Male	14.40	44.02	30.98	10.60
	Female	20.67	47.57	25.21	6.56

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 shows that for both sections of the exam, there was a significant association between the gender of the test taker and what CEFR level was obtained. However, the Cramer's V measure indicates that the effect size for both sections is very small; that is, these differences might not be sufficiently large to be meaningful.

Table 6.2.2 Overall Chi-Square Test Results for Gender and CEFR Levels

Section	χ^2	df	p	Cramer's V
Listening	312.80	3	< 0.00	0.093
Reading	494.64	3	< 0.00	0.117

On the basis of these analyses, it is not possible to say with any certainty whether male test takers are more likely to obtain higher CEFR levels on the MET than female test takers. Still, given the consistently higher number of female test takers (see Table 5.3.1), it would be important to see if these findings hold true for subsequent years of administration.

Age

Table 6.2.3 presents the overall distribution (in %) of the CEFR levels for each age band in both the listening and reading sections of the MET. The data suggests that age influences test takers' chances of obtaining higher CEFR levels. In particular, the youngest test takers (i.e., test takers who are less than 12 years old) seem to be least likely to score at the higher levels. Also, test takers in the 20–25 age group appear less likely to achieve either a C1 or B2 level. The data also suggests that, for the listening section, younger test takers (i.e. test takers who are between 13 and 20 years old) tend to score higher than older test takers (i.e., test takers who are more than 25 years old).

Table 6.2.3 Overall Percentage of Test Takers from Each Age Group Who Received Each CEFR Level for Both Sections

Section	Age	A2	B1	B2	C1
Listening	≤ 12	22.29	59.87	15.29	2.55
	13–16	14.99	40.04	31.91	13.07
	17–19	17.43	41.95	27.94	12.69
	20–22	24.83	45.78	20.56	8.83
	23–25	30.11	43.33	18.21	8.35
	26–29	26.11	42.33	21.60	9.96
	30–39	24.35	41.85	22.84	10.97
	≥ 40	32.01	37.22	21.36	9.42
Reading	≤ 12	28.03	63.69	7.01	1.27
	13–16	11.29	46.58	32.36	9.78
	17–19	11.82	46.57	32.39	9.22
	20–22	18.78	49.14	25.31	6.76
	23–25	24.28	46.37	22.19	7.16
	26–29	18.78	43.43	28.18	9.61
	30–39	14.49	43.02	31.52	10.97
	≥ 40	21.30	37.00	30.10	11.60

As in the case of gender, cross-tabulations and chi-square tests were run for each section of the exam. 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. The table shows that for both sections of the exam, there was a significant association between a test takers' age and what CEFR level was obtained. However, the Cramer's V measure again indicates that the effect size for both sections is very small; that is, these differences might not be sufficiently large to be meaningful.

Table 6.2.4 Overall Chi-Square Test Results for Age and CEFR Levels

Section	χ^2	df	p	Cramer's V
Listening	1032.30	21	< 0.00	0.094
Reading	954.97	21	< 0.00	0.091

On the basis of these analyses, it is not possible to say with any certainty whether test takers of a given age band are more likely to obtain higher (or lower) CEFR levels on the MET than others. However, if the trend of increasing age at which the MET is taken continues (see Table 5.5.2), it would be interesting to perform these analyses for subsequent administrations.

6.3. Trends in Reliability Estimates

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. For high-stakes exams such as the MET, a reliability figure of 0.80 and above is expected and acceptable.

Table 6.3.1 Reliability Ranges for the MET Listening and Reading Sections

Year	Listening	Reading and Grammar
2009	0.90–0.93	0.91–0.92
2010	0.91–0.93	0.91–0.93
2011	0.91–0.94	0.93–0.94
2012	0.91–0.94	0.92–0.94
2013	0.89–0.95	0.91–0.96

Reliability estimates are obtained for each administration of the MET. They are calculated with the program, BILOG, using the Bayes MAP (maximum a posteriori) method. Table 6.3.1 presents the lower and upper reliability estimates achieved by the tests that were administered in each year for both sections of the exam. The table shows that the reliability figures are not only above the acceptable value of 0.80, but that they are consistently at or above 0.90 each year. This suggests an excellent consistency of measurement in the MET.

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 6.4.1 SEM Estimate Ranges for the MET Listening and Reading Sections

Year	Listening	Reading and Grammar
2009	2.60–3.20	1.90–2.70
2010	2.40–3.50	1.80–3.10
2011	2.22–3.49	1.69–3.19
2012	2.16–3.16	1.69–2.78
2013	2.42–3.43	1.62–3.44

SEM estimates are obtained for each of the monthly administrations; Table 6.4.1 presents the range of SEM estimates for each year. The table shows that the SEM estimates are generally stable from year to year. Additionally, the SEM estimates as a proportion of the 80 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 speaking task is an indirect measure of the processes involved in composing an utterance, 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 MET captures evidence of a test taker's communicative language ability at the A2 – C1 levels of the CEFR through a variety of tasks in three language skills: listening, reading, and speaking. Section 2.2 described the skills and abilities expected for each language skill. Performance on the MET is expressed as a CEFR level for each test section. Reporting scores in this way is justifiable if each section can be seen to contribute differentially to the test takers' MET performance. Table 6.5.1 presents the correlation (Spearman's rho) between the listening and the reading sections for each year.

Table 6.5.1 Correlations (ρ) Between Sections[1]

Year	Correlation
2009	0.783
2010	0.810
2011	0.827
2012	0.834
2013	0.833
Overall	0.797

7. Additional MET Validity Evidence

Section 2.2 (above) presented a proposed interpretation of a test taker's MET scores. The safety of this proposed interpretation is dependent upon the evidence to support it. Test validation is the

[1] All Correlations are significant at a 0.01 level (2-tailed).

Table 7.1: Proposed Validity Claims about the MET and the Research Evidence Available

Proposed Claim	Evidence Available
The different item types and tasks are appropriate for measuring the domain and/or level of language proficiency targeted by the test.	• Aryadoust, V., & Goh, C. C. M. (2014) *Predicting listening item difficulty with language complexity measures: A comparative data mining study*, CaMLA Working Papers, 2014-2. • Chapman, M., & O'Boyle, J. (2013) *Planning time in speaking tests: How does it help?* Conference Proceedings of MexTESOL: MexTESOL 2013 Annual International Conference. Queretarro, Mexico • MacMillan, F., Chapman, M., & Stucker, J.R. (2014) A look into cross-text reading items: Purpose, development and performance. *Cambridge English: Research Notes*, *55*, 12–15. • Papageorgiou, S., Stevens, R., & Goodwin, S. (2012) The relative difficulty of dialogic and monologic input in a second-language listening comprehension test. *Language Assessment Quarterly*, *9*(4), 375–297.
Performance on the test is related to other indicators of language proficiency	• Cambridge Michigan Language Assessments [CaMLA]. (2010). Setting cut scores on the Common European Framework of Reference for the Michigan English Test: Technical Report. Ann Arbor: CaMLA.

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 MET along with the research evidence available for these claims.

7.1. The different item types and tasks are appropriate for measuring the domain and/or level of language proficiency targeted by the test.

Two recent studies investigated the design of the MET listening section. Papageorgiou, Stevens and Goodwin (2012) compared the relative difficulty of the same listening comprehension items on the MET when tested through dialogic and monologic stimuli. Previous research suggests that dialogic input may be easier to understand than monologic input because it contains more discourse markers and nonverbal cues (Fox Tree, 1999; Fox Tree & Mayer, 2008). However, Read (2002) found that monologic input may be easier to understand when compared to unscripted, spontaneous dialogue. For the current study, Papageorgiou et al. (2012) utilized three pairs of stimuli that were specifically crafted for the study. One stimulus in the pair was dialogic in nature and the other was monologic. Regardless of the input type, both contained the same content and vocabulary. Additionally, the items associated with each stimulus pair were the same, regardless of whether the stimulus was dialogic or monologic. Four items accompanied each stimulus. These stimuli and items were placed into three test forms of the MET. 494 test takers each took one of these forms of the MET (257 for form A, 138 form B and 99 form C). Rasch analyses were conducted to examine (a) the difficulty of listening comprehension items for monologic and dialogic input and to determine (b) whether the items, which were identical for the stimulus types, performed satisfactorily for both monologic and dialogic input.

The results of the Rasch analyses were mixed. For pair one, there was no difference in relative difficulty between the monologic and dialogic input types. In pair two, however, three of the four items were more difficult in connection with the monologic type. Pair three yielded a different pattern; two of the four items were more difficult in connection with the dialogic type and one was more difficult for the monologic type.

While no clear pattern emerged for the first research question, there was substantial evidence to indicate that the items performed satisfactorily regardless of the stimulus type (research question 2). Item option frequency analysis for pair one showed that test takers selected the key more frequently than any distracter for

all four items, regardless of whether the stimulus was monologic or dialogic. Additionally, correlations for the key were all positive and higher than any correlation for a distracter. This is encouraging, because more able test takers should select the key more frequently than distracters. This trend was consistent with the items associated with the stimuli in pairs two and three, as well. That is, the key was the more frequently selected option, and correlations for the key were positive and higher than any distracter. In general, the MET items analyzed in the study performed satisfactorily for both the monologic and dialogic stimuli, illustrating the soundness of MET item construction.

On the broader, theoretical point, neither stimulus type consistently appeared to be more difficult than the other. This finding is in line with previous research (Brindley & Slatyer, 2002; Read, 2002). Other factors, such as how the input is structured and lexical overlap between the stimulus and options, were found to be more predictive of item difficulty, regardless of the input type. One cannot make a claim, therefore, that an item's difficulty can be determined by the input type. There are a number of factors specific to the item itself that must be taken into consideration. Both input types, then, can accurately assess the construct of listening, and (as is the case for the MET listening section) it is important to include both in order to ensure that all aspects of the construct are being measured.

In a separate but clearly related study, Aryadoust and Goh (2014) looked at how semantic, lexical and syntactic complexity contributed to listening item difficulty on the MET. The study used seven forms of the MET. In total, 322 listening items were analyzed. There were 5,039 test takers across the seven forms. Test items were transcribed, and, using Coh-Metrix, twelve independent variables based on semantic, lexical and syntactic complexity were computed. Test item difficulty was estimated, using the Rasch model. Test items were determined to be either "low" difficulty or "high" difficulty. This was the dependent variable of the study. Four different models—regression analysis, classification and regression trees (CART), Artificial Neural Network (ANN) and Adaptive Neuro-Fuzzy Inference System (ANFIS)—were used to explore the effect of different complexity factors on MET listening items' difficulty.

Of the four models used, the ANN proved to be the most reliable. It had the highest classification accuracy, which indicates that for the 12 proposed independent variables, one could predict with relative certainty the difficulty of the item based on the presence of those variables. Knowledge of hypernyms was found to be the independent variable that influenced item difficulty the most in the ANN model. Hypernyms are semantic classes of words that are general in nature. For example, knife and fork refer to the hypernym utensil. Items that demanded a greater knowledge of hypernyms were classified as more difficult. Additionally, Flesch-Kincaid level was found to influence item difficulty nearly as much as hypernymy. These findings are important, because they add to the construct validity argument of the MET. Both hypernymy and Flesch-Kincaid level are vocabulary-related variables. Specifically, texts that are more hypernymic or have a higher Flesch-Kincaid level are cognitively more taxing due to their range in lexical features. Vocabulary acquisition studies indicate depth of vocabulary to be crucial for test takers to do well on difficult items (e.g. Buck, 2001). Similarly, another very influential variable, average givenness, which refers to the amount of overlap between the stimulus, item stems and options, has also been found to be related to item difficulty (Buck & Tatsuoka, 1998). The MET items analyzed in this study appear to be aligned with their intended difficulty, a finding made more significant by the fact that the variables contributing to item difficulty are substantiated by previous research, thus adding more evidence to the validity argument for the MET.

In a study of the reading section, MacMillan, Chapman and Stucker (2014) investigated the development and performance of cross-text reading items. The MET features four reading units comprising three to four reading passages. The passages are connected to each other by a common theme. The passages are accompanied by a number of items related to the readings. In order to answer the items correctly, test takers must process and synthesize the information from the related passages. Some of the items target information in one reading passage (regular reading items) and others can only be answered by synthesizing information from two or more passages (cross-text reading items). The development of such units is important because second language acquisition research in reading suggests that the ability to synthesize information across various sources is a skill that separates higher-level learners from lower-level learners (e.g. Jamieson, Jones, Kirsch, Mosenthal & Taylor, 2000). Cross-text reading items also respond to an identified need for K-12 students to be able to integrate information across multiple sources (cf Gere, 2009; Rosenfeld, Leung, and Oltman, 2001).

This study analyzed 200 items (100 cross-text items and 100 regular reading items) from recently administered MET and ECCE exams. The cross-text items required test takers to use information from two or more texts to answer the question. It was found that both item types achieved similar discrimination results; 94% of the cross-text items and 93% of the regular items met CaMLA's criteria for acceptable discrimination. However, the items related to the cross-text item type were more likely to return facility values that fell within the specifications established for the test; 99% of cross-text items yielded acceptable facility values whereas 88% of the regular items yielded such facility values. This study provides evidence that while both item types performed well, more cross-text reading items achieved acceptable facility values. The findings of this study, then, help to contribute to the construct validity argument for including these types of items on the MET.

As part of the development of the MET speaking test, Chapman and O'Boyle (2013) explored whether planning time benefits test takers during speaking tests. The MET speaking test has five tasks which progress in difficulty; the fourth and fifth tasks are considered to be the most cognitively demanding. Previous research concerning the benefits of planning time is mixed. In the classroom, planning time has been found to be beneficial but its benefits have varied from study to study. Crookes (1989) found that when learners were given planning time they produced a wider variety of vocabulary but were not more accurate. Foster and Skehan (1996), however, found that planning time resulted in more accuracy but did not influence the complexity of language. In the context of a language test, planning time has not been found to affect test scores (Wigglesworth, 1997) nor does it affect the test takers' response characteristics, i.e. accuracy, fluency, complexity (Iwashita, McNamara, and Elder, 2001).

This study specifically focused on test takers' desire for planning time when answering tasks on the MET speaking test. Twenty-two English language learners volunteered to take one of two forms of MET speaking tests. Twenty-five tests were administered. Three test takers took the test in both forms (the first with planning time, the second without). Twelve of the participants were allowed 30 seconds to prepare their responses, while 13 of the participants were given no planning time. Participants were given a questionnaire following the test. The majority of the participants who received no planning time indicated that they had sufficient time to answer their prompts. Only two participants who were given no planning time indicated on Task 4 that they did not have enough time, and only three participants indicated the same on Task 5. Additionally, it was observed that participants who were given planning time did not use their allotted time.

These findings provide support for the decision to exclude planning time from the MET speaking test. The majority of participants in the study, whether in the planning time or no planning time condition, felt that they had sufficient time to give their responses. Additionally, the majority of those who were given planning time did not fully use it. Research also suggests that while planning time may provide for a higher quantity of output, it does not result in higher scores (Wigglesworth, 1997). The design of the MET speaking test, then, is appropriate based on the literature and the results of this study.

7.2. Performance on the test is related to other indicators of language proficiency

The current CEFR cut scores for the MET were established empirically by a standard-setting panel comprised of 13 judges, all of whom were teachers, teacher trainers, academic directors and academic advisors at centers that administered the MET in Colombia. A technical report detailing the study is available on the CaMLA website (CaMLA, 2010).

According to the Council of Europe's Manual (2009), standard-setting meetings must address procedural, internal and external validity. The standard-setting meeting proceeded through three clear stages: familiarization (in which the judges were familiarized with the CEFR); training (in which the judges became comfortable with the judgment task); and judgment (in which the judges set cut scores for each CEFR level). At the end of the process, the panelists completed a feedback questionnaire. The results of this questionnaire established procedural validity; all participants understood their tasks and were confident in the judgments that they had made.

The judgment task used a modified Angoff (1971) approach. Analysis of the data showed that the judges were quite accurate in their ability to match descriptors to the appropriate CEFR level; Spearman correlations ranged 0.71–0.97 ($p \geq 0.01$) across the different language skills/elements of reading, listening, grammar, vocabulary. Additionally, judges were consistent with each other in their matching of

level descriptors; Cronbach's alpha was stable at 0.98 ($p \geq 0.01$) across the language skills/elements. When establishing cut scores for each section, intra- and inter-judge consistency was high (ranging 0.83–0.94, $p \geq 0.01$). The resulting cut scores for the listening and the reading/grammar sections demonstrated decision consistency. These results established internal validity of the standard setting procedure.

To establish external validity the panel reviewed the proposed cut scores in relation to how these cut scores would classify 660 MET test takers. Using these cut scores, the majority of the test takers would be at the B1/B2 levels; all judges felt that this was reasonable based on their knowledge of the MET and its test takers (CaMLA, 2010: 11). Overall, this report of the MET–CEFR linking study provides good evidence of the claims that CaMLA makes about the proficiency levels tested by the exam.

The research already completed has begun the work of building a validity argument for the MET. 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 MET:

- The structure of the test is consistent with its stated construct and with the way in which scores are reported.
- The listening and reading tasks elicit language knowledge and processes expected for the domain and/or level of language proficiency targeted by the test.
- The language elicited by the speaking section reflects the domain and/or level of language expected.
- The rating scale for the speaking section appropriately distinguishes between test takers of different levels of language proficiency.
- MET test results are used appropriately.
- The MET has positive consequences for stakeholders.

8. References

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

Angoff, W. H. (1971) Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–6000). Washington: American Council on Education.

Aryadoust, V., & Goh, C. C. M. (2014). Predicting listening item difficulty with language complexity measures: A comparative data mining study, CaMLA Working Papers, 2014–2, Ann Arbor, MI: CaMLA.

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

Brindley, G., & Slatyer, H. (2002) Exploring task difficulty in ESL listening assessment. *Language Testing, 19*, 369–394.

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

Buck, G., & Tatsuoka, K. (1998) Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15*(2), 119–157.

Cambridge Michigan Language Assessments. (2010). Setting cut scores on the Common European Framework of Reference for the Michigan English Test: Technical Report. Ann Arbor: CaMLA.

Chapman, M., & O'Boyle, J. (2013) *Planning time in speaking tests: How does it help?* Conference Proceedings of MexTESOL: MexTESOL 2013 Annual International Conference. Queretarro, Mexico

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

Council of Europe (2009) *Relating language examiniations to the Common European Framework of Reference for Languages: Learning, teaching, assessment. A Manual*. Retrieved from http://www.coe.int/t/dg4/linguistic/Source/ManualRevision-proofread-FINAL_en.pdf

Crookes, G. (1989) Planning and interlanguage variation. *Studies in Second Language Acquisition, 11*, 367–383.

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.
Foster, P., & Skehan, P. (1996) The influence of planning and task type on second language performance. *Studies in Second Language Acquisition 18*, 299–323.
Fox Tree, J. E. (1999). Listening in on monologues and dialogues. *Discourse Processes*, 27, 35–53.
Fox Tree, J. E., & Mayer, S. A. (2008). Overhearing single and multiple perspectives. *Discourse Processes*, 45, 160–179.
Gere, A. R. (2009) Literacy learning in the 21st century: A policy brief produced by the National Council of Teachers of English, *The Council Chronicle, 18*(3), 14–16.
Iwashita, N., McNamara, T., & Elder, C. 2001. Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning*; vol. 51:3, 401–436.
Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (2000) *TOEFL 2000 Framework: A Working Paper*, TOEFL monograph MS-16, Princeton: Educational Testing Service.
MacMillan, F., Chapman, M., & Stucker, J.R. (2014) A look into cross-text reading items: Purpose, development and performance. *Cambridge English: Research Notes, 55*, 12–15.
Papageorgiou, S., Stevens, R., & Goodwin, S. (2012). The relative difficulty of dialogic and monologic input in a second-language listening comprehension test. *Language Assessment Quarterly*, 9 (4), 375–297.
Read, J. (2002). The use of interactive input in EAP listening assessment. *Journal of English for Academic Purposes*, 1, 105–119.
Read, J. (2002) The use of interactive input in EAP listening assessment. *Journal of English for Academic Purposes*, 1, 105–119.
Rosenfeld, M., Leung, S., and Oltman, P. (2001) *The reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels*, TOEFL monograph MS-21, Princeton: Educational Testing Service.
Wigglesworth, G. (1997) An investigation of planning time and proficiency level on oral test discourse. *Language Testing, 14*(1), 85–106.