2011–2014
Technical Review

Contact Information

All correspondence and mailings should be addressed to:

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

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

info@cambridgemichigan.org
CambridgeMichigan.org

07/2015

Table of Contents

List of Tables

1. Introduction

The Michigan English Language Assessment Battery (MELAB) is a test of English language proficiency for intermediate and advanced learners of English. In the period 2011 to 2014, the exam was administered monthly at test centers around the world.

This report provides test users with technical information about the MELAB. Section 2 provides general information about the test and a proposed interpretation of MELAB 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 and decision making. Section 4 describes the changes in the MELAB from 2011 to 2014. Section 5 discusses the MELAB test-taking population, looking particularly at the yearly distributions of test takers by gender, age, and purpose for taking the test. Section 6 looks at trends in the MELAB test results by section and demographic variables. It also examines trends in reliability estimates, standard error of measurement, and subtest correlations for each year. The final section of the report reviews the validity evidence currently available to support CaMLA's proposed interpretation of the MELAB results.

2. Description of the MELAB

2.1 General Description

The MELAB is a standardized, English as a foreign language examination for adult nonnative speakers of English who will need to use English for academic or professional purposes. The four component skills of listening, reading, writing, and speaking are evaluated through a combination of tasks.

The MELAB is aimed at the B1–C1 levels of the Common European Framework of Reference (CEFR; Council of Europe, 2001), and the score report is valid for two years. MELAB scores are used by students applying to universities where the language of instruction is English. They are also used for certification of English proficiency for various organizations and licensing professionals.

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 confirm 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 MELAB is open to all people who wish to take the exam.

2.2 Proposed Interpretation of Scores

The MELAB is intended for adult nonnative speakers of English who are seeking admission to colleges and universities where the language of instruction is English, or for professional purposes. Consequently, the content and tasks are drawn from the formal and informal communication contexts a college or university student might encounter, as well as general occupational or office settings. These include conversations between friends and service encounters as well as the interactions and inputs that might be expected in seminars and lectures.

The MELAB is a multilevel exam, covering a range of proficiency levels on the CEFR (Council of Europe, 2001) from B1 to C1; test takers at the B1 and B2 levels are considered independent users of English, and test takers at the C1 level are considered proficient users of English. Selected CEFR performance descriptors illustrating what candidates should be able to do at each level are available on the CaMLA website.

2.3 Test Structure

The MELAB measures four skill areas: listening, reading, writing, and speaking. The writing, listening, and reading sections are taken in one sitting. The speaking test is optional and is taken separately. Table 2.3 describes the format and content of the MELAB. Test preparation resources are available on the CaMLA website.

3. Scoring and Reporting of MELAB Results

3.1 Explanation of Scoring for Each Section

The MELAB speaking and writing sections are graded according to scales established by CaMLA (see the website for the rating scales). The speaking section is conducted and assessed by a CaMLA-certified speaking examiner, and the writing section is assessed by at least two CaMLA-certified raters.

Table 2.3: Format and Content of the MELAB

Section	Time	Description		
Writing	30 minutes	Test takers write an essay based on one of two topic choices.	1 task	
Listening	35–40 minutes	**Part 1** A short recorded question or statement is accompanied by three printed responses. Test takers choose the statement that conveys a reasonable answer or response.	60	Multiple-Choice Items
		Part 2 A recorded conversation is accompanied by three printed statements. Test takers choose the statement that means about the same thing as what is heard.		
		Part 3 Four recorded interviews, such as those that might be heard on the radio, are each followed by recorded comprehension questions. The questions and answer choices are printed in the test booklet. Test takers choose the correct answer from the choices.		
Grammar Cloze Vocabulary Reading (GCVR)	80 minutes	**Grammar** An incomplete sentence is followed by a choice of four words or phrases to complete it. Only one choice is grammatically correct.	110	
		Cloze Two passages with deletions are followed by choices of words and phrases to complete the text. Test takers must choose the word or phrase that best fills the blank in terms of grammar and meaning.		
		Vocabulary An incomplete sentence is followed by a choice of four words or phrases to complete it. Test takers must choose the option that best completes the sentence in terms of meaning.		
		Reading Four reading passages are followed by comprehension questions. Test takers choose the correct answer from the printed answer choices.		
Speaking	15 minutes	Test takers engage in a conversation with an examiner.		

The listening and grammar, cloze, vocabulary, and reading (GCVR) sections are scored electronically at CaMLA. Each correct answer contributes proportionally within each section, and no points are deducted for wrong answers. A scaled score is calculated using an advanced mathematical model based on Item Response Theory. This method ensures that scores are comparable across different administrations.

3.2 Equating Procedures

In order to ensure that the MELAB 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. A proportion of items on each exam are designated as the common (link) items that are used to equate the different exam forms using item difficulty. Item difficulties from previous administrations are stored in a database. When 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 MELAB to be equated. The scale and location parameters are computed separately for each section of the exam.

3.3 Procedures for Reporting Scores

MELAB scores are reported on an official score report form. Official score reports are sent directly from CaMLA to the institutions or admissions offices indicated by the test taker. The score report provides the following information:

- A score for each of the sections
- The final MELAB score, which is the average of the scores for the writing, listening, and GCVR sections
- A speaking test score for test takers who opted to take this part of the test
- Additional comments about the test performance (where this is relevant)

Table 3.3 presents the score ranges for each MELAB section.

Table 3.3: MELAB Score Ranges

Section	Range	Notes
Writing	0–97	
Listening	0–100	
GCVR	0–100	
Speaking	1–4	May include + or - in the scores
Final MELAB Score	0–99	Average of writing, listening, and GCVR scores

3.4 Interpretation of Scores for Each Section

As stated in the description of the exam (Section 2.2), the MELAB covers a range of proficiency levels (B1–C1) on the CEFR. In general, test takers who receive scores at the 50th percentile in each section can be expected to have the following skills and abilities:

Speaking: They are quite fluent and their accent is usually intelligible even where there are deviations from conventional pronunciation. They are active in conversation and can elaborate on topics. They have a good vocabulary but there are gaps in their linguistic range and control.

Writing: They are able to develop on an assigned topic but do not cover all aspects of the issue. Organization is acceptable but they do not always connect their ideas well. Vocabulary is adequate but there are gaps in their linguistic range and control.

Listening: They are able to understand the main idea in conversation and discussion on topics they regularly encounter. They are comfortable listening to speech in formal contexts such as lecture presentations; in informal contexts they benefit from opportunities to seek clarification.

Reading: They are able to understand written materials on topics that they regularly encounter. They are able to identify the main idea of a text, and can locate important details as well as infer attitudes or feelings.

Use of English: They are able to communicate with reasonable accuracy in familiar contexts. Errors occur but it is clear what they are trying to express. Their vocabulary is sufficient to express themselves on familiar topics.

3.5 Guidelines for Decision Making

When interpreting a MELAB score report, it is important to remember that the MELAB estimates the test takers' true proficiency by approximating the kinds of tasks that they may encounter in real life. Also, temporary factors unrelated to an examinee's proficiency, such as fatigue, anxiety, or illness, may affect exam results.

When using test scores for decision making, users should check the date the test was taken. While the score 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 to use English on a regular basis. Additionally, both section scores and the final MELAB score should be considered. Two test takers who have the same final scores but quite different section scores may differ in their language skills. Such differences may affect their ability to use English effectively in different contexts.

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 context of academic studies in English, the ability to function effectively involves not only knowledge of English, but also other knowledge and skills such as intellectual knowledge and study skills. Since language ability is just one of many factors that affect success, or lack of success, in an academic setting, MELAB scores should not be used to predict academic success or failure.

4. Changes to the MELAB from 2011–2014

During the period covered by this report CaMLA introduced three changes to the MELAB. One was a minor revision to the layout of the test and two were changes to the design of the test.

4.1 Changes to the Layout of the Test

Writing Test Booklet (November 2012)

Until November 2012, both of the writing prompt options were printed on the front of the MELAB writing test booklet. Test taker feedback revealed that this was inconvenient and required flipping back and forth from their essay to the cover to remind themselves of specific details of the question. In order to improve the test-taking experience the writing prompts now appear on the inside cover of the booklets, just above the lines where test takers begin writing their essay.

4.2 Changes to the Design of the Test

Two Cloze Passages (November 2011)

Until November 2011, the MELAB cloze section featured one long cloze passage with twenty items. The MELAB cloze section now comprises two shorter cloze passages, giving test takers a greater variety of topics to engage with. The first passage contains ten items and the second passage contains fourteen items. The grammar and vocabulary subsections have been shortened to accommodate the additional cloze items. The total number of GCVR items to be answered has not changed.

Printing Question Stems in Test Booklets for Listening Interview Items (July 2012)

Until July 2012, 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 a clearer purpose while listening, a small change was made 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 examinees to read the questions while they are listening to the radio report.

5. MELAB Test-Taking Population

This section presents an overview of the test takers who took the MELAB during the period covered by this report, providing demographic information for the testing population. Every MELAB test taker completes a registration form, which asks for gender, date of birth, first language, and purpose for taking the test. Cases were information has not been provided or has not been correctly given are treated as missing data.

5.1 First Language

The MELAB attracts test takers from a wide range of first language backgrounds; test takers from 96 first language backgrounds took the MELAB in the period 2011 to 2014. Table 5.1 lists the ten largest MELAB first language backgrounds.

Table 5.1: Ten Largest MELAB First-Language Backgrounds

Arabic	Punjabi
Chinese (Cantonese/Mandarin)	Russian
Farsi/Persian	Spanish
Korean	Tagalog/Filipino
Malayalam	Urdu

5.2 Gender Distribution

Table 5.2 presents the distribution of test takers by gender. It shows that during the period covered by this report, the percentage of male test takers has increased, while the percentage of female test takers has decreased. Specifically, in 2011 and 2012 female test takers

accounted for nearly two thirds of the test population while in 2013 and 2014 male test takers accounted for just over half of the test population.

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

Gender	2011	2012	2013	2014
Male	33.82	32.59	51.29	54.87
Female	65.59	61.24	48.18	43.78
Missing Data	0.59	6.18	0.54	1.35

5.3 Age Distribution

Table 5.3 presents the distribution of test takers by age. It indicates that the MELAB population is generally trending toward younger test takers. Since 2013, the proportion of test takers over 30 has decreased, while the proportion of test takers between 17 and 22 has increased. Additionally, the average age of the test takers has fallen each year. The average test taker age in 2011 was 31.44 years and in 2014 it was 26.84 years.

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

Age	2011	2012	2013	2014
13–16	0.27	0.47	0.90	0.28
17–19	10.73	13.98	22.71	22.89
20–22	9.41	10.15	17.87	22.25
23–25	13.00	12.03	13.45	13.86
26–29	15.05	15.14	12.91	12.58
30–39	30.23	29.48	19.73	14.85
≥ 40	21.27	18.53	12.01	12.72
Missing Data	0.05	0.22	0.42	0.57

5.4 Distribution by Purpose for Taking the Test

Table 5.4 presents the distribution of test takers by purpose for taking the MELAB. It shows that the main purpose for taking the MELAB has changed during this four year time period. In 2011 and 2012 the majority of the test takers took the test for professional purposes, and in 2013 and 2014 the majority took it for education purposes.

Table 5.4: Distribution (in %) of MELAB Test Takers by Purpose for Taking the Test

Purpose	2011	2012	2013	2014
Educational	37.91	30.38	67.30	80.53
Professional	57.09	52.28	17.27	8.32
Other	4.27	4.91	12.37	10.16
Missing Data	0.73	12.43	3.05	1.00

5.5 Cross-Tabulation Analysis

We explored patterns in the demographic data (5.2–5.4) using cross tabulations. It should be noted that for the analysis presented in this section that involves the age variable, the two youngest age bands were combined due to the small percentage of test takers in the 13–16 year age band (see Table 5.3).

Gender and Age

Table 5.5.1 presents the distribution of test takers by age band for each gender. It shows that in general, the female test takers were older than the male test takers. However, the table also shows that the male and female age distributions were more similar in 2013 and 2014 than they were in 2011 and 2012. This suggests that while there may be a relationship between gender and age, the test population has changed in such a way that the strength of the association has weakened.

The Pearson chi-squared (χ^2) test of independence was applied to determine whether or not there was a statistically significant relationship between test taker gender and age. This statistical test examined the null hypothesis of independence against the alternative of dependence to determine whether or not knowledge of one variable could help to predict the other. The χ^2 test of independence was selected because it allows for the comparison of two categorical variables. A measure of effect size, Cramer's V, was also used. It provided a measure of the strength (meaningfulness) of the association between two variables, taking account of sample size and degrees of freedom (Field, 2005: 692).

Table 5.5.2 summarizes the χ^2 value, the degrees of freedom (df), the level of significance (p), and Cramer's V. It shows that there was a significant

Table 5.5.1: Distribution (in %) of Test Takers by Age for Each Gender

Year	Gender	≤ 19	20–22	23–25	26–29	30–39	≥ 40
2011	Female	7.98	6.45	11.93	16.09	35.78	21.78
	Male	16.94	15.05	15.32	13.04	19.49	20.16
2012	Female	10.63	5.79	10.21	16.41	35.77	21.19
	Male	21.80	17.35	14.79	13.46	18.13	14.46
2013	Female	22.79	14.45	13.95	13.20	22.29	13.33
	Male	24.77	21.38	12.97	12.85	17.52	10.51
2014	Female	21.51	16.42	17.57	14.61	16.91	12.97
	Male	24.77	27.37	11.15	11.02	13.36	12.32

association between gender and age for each year. Additionally, Cramer's V indicates moderate association between gender and age in 2011 and 2012, and a weak association in 2013 and 2014. These results confirm that the strength of the relationship between test taker gender and age has decreased over the period covered by this report.

Table 5.5.2: Chi-Square Test Results for Gender and Age

Year	χ^2	df	*p*	Cramer's V
2011	125.54	5	< 0.001	0.240
2012	220.34	5	< 0.001	0.292
2013	19.57	5	0.003	0.109
2014	36.13	5	< 0.001	0.162

Gender and Purpose for Taking the Test

Table 5.5.3 presents the distribution of test takers by purpose for taking the test for each gender. It shows that in 2011 and 2012, the majority of female test takers took the test for professional purposes, and the majority of male test takers took the test for educational purposes. However, in 2013 and 2014, it shows that the majority of both male and female test takers took the test for educational purposes. This suggests that there is a relationship between gender and purpose for taking the MELAB, and that this relationship has changed over time.

The Pearson χ^2 test of independence was used to determine whether or not there was a statistically significant relationship between test taker gender and purpose for taking the test. Table 5.5.4 summarizes the χ^2 value, the degrees of freedom (df), the level of significance (*p*), and Cramer's V. It shows that there was a significant association between gender and purpose for taking the MELAB from 2011 to 2013, but that there was no significant association between them in

Table 5.5.3: Distribution (in %) of MELAB Test Takers by Purpose for Taking the Test for Each Gender

Year	Gender	Educational	Professional	Other
2011	Female	27.97	67.11	4.93
	Male	58.01	38.90	3.10
2012	Female	23.65	69.86	6.49
	Male	58.42	37.82	3.76
2013	Female	66.04	19.92	14.04
	Male	72.72	15.71	11.57
2014	Female	80.82	9.02	10.16
	Male	81.51	8.07	10.42

2014. Additionally, Cramer's V indicates a moderate association between gender and purpose for taking the test in 2011 and 2012, and a negligible association in 2013 and 2014. These results indicate that the relationship between gender and purpose for taking the test has diminished over time.

Table 5.5.4: Chi-Square Test Results for Gender and Purpose for Taking the Test

Year	χ^2	df	*p*	Cramer's V
2011	187.52	2	< 0.001	0.293
2012	280.46	2	< 0.001	0.340
2013	8.56	2	0.014	0.073
2014	0.40	2	0.820	0.017

Age and Purpose for Taking the Test

Table 5.5.5 presents the distribution of test takers by age band for each purpose for taking the test. It shows that in general, test takers who took the MELAB for educational purposes were younger than those who took it for professional or other purposes. This is to be expected since younger test takers are more likely to be in school or seeking entry to a university, while older test takers are more likely to be employed or seeking employment.

The Pearson χ^2 test of independence was used to determine whether or not there was a statistically significant relationship between test taker age and purpose for taking the test. Table 5.5.6 summarizes the χ^2 value, the degrees of freedom (df), the level of significance (*p*), and Cramer's V. It shows that there was a significant association between age and purpose for taking the test for each year. Additionally, Cramer's V indicates a moderate to strong association between age and purpose for taking the test for each year. This indicates a persistent relationship between age and purpose for taking the test.

Table 5.5.6: Chi-Square Test Results for Age and Purpose for Taking the MELAB

Year	χ^2	df	*p*	Cramer's V
2011	719.32	10	< 0.001	0.406
2012	1053.20	10	< 0.001	0.466
2013	517.93	10	< 0.001	0.400
2014	335.33	10	< 0.001	0.348

General Trend

The results of the analyses presented in this section suggest that though the composition of the test population appears to have altered over the period covered by this report, the MELAB test population is primarily comprised of two distinct subgroups.

Table 5.5.5: Distribution (in %) of MELAB Test Takers by Age for Each Purpose for Taking the Test

Year	Purpose	≤ 19	20–22	23–25	26–29	30–39	≥ 40
2011	Educational	27.82	18.71	16.07	15.23	14.99	7.19
	Professional	0.40	3.26	10.83	14.89	40.53	30.10
	Other	4.30	7.53	17.20	15.05	27.96	27.96
2012	Educational	36.67	19.05	15.48	11.79	12.26	4.76
	Professional	0.28	2.42	10.16	17.97	41.74	27.44
	Other	2.94	4.41	10.29	18.38	33.09	30.88
2013	Educational	32.62	22.82	15.33	12.39	12.66	4.19
	Professional	1.04	2.08	6.60	13.19	38.89	38.19
	Other	8.74	12.62	12.14	16.50	32.04	17.96
2014	Educational	27.33	26.00	14.64	12.95	13.13	5.94
	Professional	0.00	2.56	8.55	10.26	28.21	50.43
	Other	10.64	9.93	12.77	10.64	19.86	36.17

One group takes the test for educational purposes, is generally younger (≤ 25), and the test takers are more likely to be male. The other group takes the test for professional purposes, is generally older (> 25), and the test takers are more likely to be female. Correspondence with test centers confirms this pattern and also indicates that the group that takes the MELAB for professional purposes have often been living and working in an English-medium context for a number of years before taking the test. On the other hand, the test takers who take the MELAB for educational purposes are usually newly arrived to the English-medium context in which they take the test. This opens up the possibility that the two groups could have different performance profiles on the MELAB (see Section 6.2).

6. MELAB Results and Test Statistics

6.1 Trends in Descriptive Statistics for Final MELAB Score and Individual Sections

The MELAB score report provides test takers' scaled scores for each written section (listening, GCVR, and writing) as well as a final score that is an average of the scaled scores for the written portion of the test. The speaking test result is reported separately. This subsection begins by looking at the descriptive statistics for the final score as well as individual sections. Table 6.1.1 shows the descriptive statistics for the MELAB final scores. It indicates that there is a downward trend in test takers' final scores. During the period covered by this report, the average final score decreased by just over two points, and the median final score decreased by three points. The standard deviation also increased in 2013 and 2014, which suggests that the test takers' MELAB scores were more varied than in the previous two years. Overall, the information in Table 6.1.1 suggests that the MELAB test population has become slightly less proficient during the period covered by this report. It is important to note, however, that these differences are relatively small as a proportion of the 100-point scale. Potential sources of this trend can be more closely examined by looking for trends in each section of the MELAB.

Statistic	2011	2012	2013	2014
Minimum Score	45	44	44	32
Maximum Score	98	97	98	98
Median Score	77	76	75	74
Average Score	76.17	75.13	74.70	73.75
Standard Deviation	9.72	9.97	11.62	11.64

Listening

Table 6.1.2 shows the descriptive statistics for the MELAB listening section scores. It reveals a slight downward trend in the section's average score. During the period covered by this report, both the average and median listening scores decreased by two points. The standard deviation also slightly increased in 2013 and 2014, which suggests that the test takers' listening scores were more varied than in the previous two years. Overall, the information in Table 6.1.2 suggests that the MELAB test population has become slightly less proficient on the listening section during the period covered by this report. It is important to note, however, that these differences are relatively small as a proportion of the 100-point scale.

Table 6.1.2: Descriptive Statistics for MELAB Listening Section Scores

Statistic	2011	2012	2013	2014
Minimum Score	33	34	31	0
Maximum Score	100	98	98	98
Median Score	79	78	78	77
Average Score	77.46	76.34	76.10	75.33
Standard Deviation	12.29	12.43	13.28	13.43

GCVR

Table 6.1.3 shows the descriptive statistics for the MELAB GCVR section scores. It reveals a downward trend in the section's average score. During the time period covered by this report both the average and median GCVR scores decreased by four points. The standard deviation also increased in 2013 and 2014, which suggests that the test takers' GCVR scores were more varied than in the previous two years. Overall, the information in Table 6.1.3 suggests that the MELAB test population has become slightly less proficient on the GCVR section during the period covered by this

report. As is the case for the listening section, these differences are relatively small as a proportion of the 100-point scale.

Table 6.1.3: Descriptive Statistics for MELAB GCVR Section Scores

Statistic	2011	2012	2013	2014
Minimum Score	33	22	29	22
Maximum Score	100	98	99	99
Median Score	78	76	74	74
Average Score	75.77	73.84	72.15	71.95
Standard Deviation	13.24	14.75	16.10	16.17

Writing

Table 6.1.4 shows the descriptive statistics for the MELAB writing section scores. It indicates that performances on this section were relatively stable over time. The only notable trend is the increase in the standard deviation figures in 2013 and 2014 suggesting (like the figures for listening and GCVR) that the test takers were more varied in their language proficiency.

Table 6.1.4: Descriptive Statistics for MELAB Writing Section Scores

Statistic	2011	2012	2013	2014
Minimum Score	53	53	53	53
Maximum Score	97	97	97	97
Median Score	75	75	75	75
Average Score	75.28	75.19	75.97	74.36
Standard Deviation	7.26	6.43	8.46	8.57

Speaking

Table 6.1.5 shows the distribution of MELAB speaking test scores. It indicates that the distribution of scores was relatively consistent over time. The data also indicates that a large number of test takers receive the highest scores for the speaking test (4- and 4). This is out of line with the trends for the written section of the exam and suggests that many of the test takers who opted for the speaking test were highly proficient English speakers. However, this pattern might also suggest a need for additional training for MELAB speaking test examiners. CaMLA's quality assurance team 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 remains an ongoing task.

Table 6.1.5: Distribution (in %) of MELAB Speaking Section Scores

Score	2011	2012	2013	2014
1	0.00	0.00	0.35	0.00
1+	0.20	0.00	0.58	0.34
2-	0.34	0.17	1.27	1.37
2	0.87	0.44	1.50	2.92
2+	2.62	2.50	4.72	5.50
3-	6.58	8.55	7.71	7.73
3	16.32	19.32	16.23	13.75
3+	24.58	26.82	20.37	21.31
4-	27.13	26.93	25.09	25.77
4	21.36	15.27	22.21	21.31

6.2 Analysis of Results by Gender, Age, and Purpose for Taking the Test

This subsection reflects upon the relationship between key demographic characteristics and the MELAB final scores for the overall test population examined in the time frame covered in this report. The averages scores were obtained for the different groups of each demographic variable. The results were then closely examined to determine whether or not group membership had a statistically significant effect on the average final scores.

Gender

Table 6.2.1 presents the average MELAB final scores for test takers by gender. It indicates that female test takers tend to score higher than male test takers. A Welch two sample t-test was performed to determine whether or not the difference between male and female average scores was statistically significant. This statistical test allows the means of two continuous variables to be compared even when the variances are unequal. It examines the null hypothesis of equal means against

the alternative of unequal means. A measure of effect size, Cohen's *d*, was also used. It is a measure of the standardized difference between means for two groups.

Table 6.2.1: Average MELAB Final Scores by Gender

Gender	Average Score
Male	73.86
Female	76.12

The two sample t-test showed that there was a statistically significant difference between the average scores of male and female test takers ($t = 9.08$, $df = 6141$, $p < 0.001$). Cohen's *d* showed that there was a small effect size ($d = 0.22$). These results provide evidence that there is a small but meaningful relationship between test taker gender and the average MELAB final score. In general, female test takers appear to be more likely to score well on the MELAB than male test takers.

Age

Table 6.2.2 presents the average MELAB final scores for test takers by age band. It is important to note that the percentage of the test-taking population in the youngest age group is very small (see Table 5.3, above) so the average score for this group should be interpreted with caution. Nevertheless, the data in Table 6.2.2 suggests that younger test takers (i.e., ≤ 25 years old and probably still in education) tend to perform less well on the MELAB than older test takers. In order to establish whether the effect of age was meaningful, we performed a simple linear regression, using age as the predictor variable and MELAB final score as the outcome variable. Because simple linear regression calculations demand continuous variables, in this calculation we did not use the age bands reported in Table 6.2.2. Instead the test takers' age at the time of taking the test was used as the predictor variable. A measure of effect size, the coefficient of determination (R^2), was also used to measure the degree of association between the variables. When multiplied by 100, this measure can be interpreted as the percent of the outcome variance explained by the regression model.

Table 6.2.2: Average MELAB Final Scores by Age Group

Age	Average Score
13–16	72.05
17–19	73.06
20–22	71.66
23–25	73.24
26–29	75.57
30–39	76.83
≥ 40	78.34

The results of the regression analysis indicate that there was a statistically significant relationship between test taker age and MELAB final score ($F = 80.2$, $df_1 = 1$, $df_2 = 8024$, $p < 0.001$). Analysis of the regression equation (Score = 0.22*Age + 68.61) reveals that there is a positive linear relationship between age and final score. This means that as test taker age increases, test taker final score also tends to increase. While the effect of a single additional year of age on a test taker's final MELAB score is rather small (0.22 of a point on the scale), the effect of multiple additional years of age has a larger effect on the test taker's final MELAB score. Additionally, the coefficient of determination showed that while there was a small effect size ($R^2 = 0.042$), the regression model only explained 4.20% of the variability in final score.

In order to obtain a better understanding of the effect of test taker age on the MELAB final score, the data was divided into two groups: test takers who are 25 years of age or younger (average score = 72.66) and test takers who are over 25 years of age (average score = 76.96). These two subgroups were selected because they have similar average final scores (see Table 6.2.2), and because they correspond to the age bands of the two distinct groups of test takers highlighted in Section 5.5. A Welch two sample t-test was used to analyze this data to compare the average scores of the two variables. The t-test showed that there was a statistically significant difference between test taker scores ($t = 17.91$, $df = 6639$, $p < 0.001$), and Cohen's *d* showed a moderate effect size ($d = 0.41$). These results show that that there is a meaningful relationship between test taker age and average MELAB final score. In general, older test takers (i.e., > 25) are more likely to receive higher scores on the MELAB than younger test takers (i.e., ≤ 25).

Purpose for Taking the Test

Table 6.2.3 presents the average MELAB final scores for test takers by purpose for taking the test. It suggests that test takers who take the MELAB for professional purposes tend to score higher than test takers who take it for educational purposes. A Welch two sample t-test was performed to determine whether or not the difference between educational and professional purpose average scores was statistically significant.

Table 6.2.3: Average MELAB Final Scores by Purpose for Taking the Test

Purpose	Average Score
Educational	72.46
Professional	78.75

The Welch two sample t-test showed that there was a statistically significant difference between the average scores of test takers who took the MELAB for educational and professional purposes ($t = 27.25$, $df = 6958$, $p < 0.001$). Cohen's d showed that there was a moderate effect size ($d = 0.63$). These results provide evidence that there is a meaningful relationship between purpose for taking the test and the average MELAB final score. In general, test takers who take the test for professional purposes appear to be more likely to score well on the MELAB than test takers who take the test for educational purposes.

General Trend

Overall, the analysis of test taker performance by gender, age, and purpose for taking the test has revealed systematic differences in the performances that correspond to the two subgroups identified in Section 5.5. That is, there is a group of test takers who take the MELAB for educational purposes and who are typically under 25 years of age, and another group of test takers who take the MELAB for professional purposes and who are typically over 25 years of age. Our analyses suggest that the group of test takers who take the test for professional purposes are more likely to perform well on the MELAB. Given that this group of test takers has often been living and working in an English-medium context for a number of years before taking the test, this pattern of results is unsurprising. Nevertheless, it would be interesting to confirm through differential item functioning (DIF) analyses that the MELAB offers all test takers equal opportunity to perform well on the exam.

6.3 Trends in Reliability Estimates, Standard Error of Measurement, and Rater Agreement Statistics

Test scores are a numerical measure of a test taker's ability. *Reliability* refers to the consistency of the 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 administration to another due to variation in the items across different forms of the same test or due to variability in individual performance. Among the reasons for this are temporary factors unrelated to a test taker's proficiency, such as fatigue, anxiety, or illness. As a result, test scores always contain a small amount of measurement error. The aim, however, is to keep this error to a minimum. For high-stakes English language proficiency tests such as the MELAB, a reliability figure of 0.80 and above is expected and acceptable.

Apart from monitoring reliability, the estimated variability in test taker performance is also monitored through the standard error of measurement (SEM) estimate. Test scores always contain a small amount of measurement error. This can be monitored by calculating the SEM. About two-thirds of the time, a test taker's score should be expected to fall in the interval of 1 SEM unit around his or her test score. The smaller the SEM, the narrower the size of this interval around the test score.

Reliability and SEM estimates are obtained for each administration of the MELAB. The reliability estimates are calculated in Excel using the KR-20 (Kuder-Richardson Formula 20) method. The SEM estimates are calculated using the reliability estimates and the scaled scores and are reported in terms of the 100-point scaled score that is used for score reporting. Table 6.3.1 lists the average reliability and SEM estimates for each year. It shows that both estimates are generally stable from year to year. The table also shows that the reliability estimate for the listening section is typically lower than that of the GCVR section. This is because of the relative length of the sections; the listening section comprises 60 items whereas the GCVR section comprises 110 items. Nevertheless, the reliability

Table 6.3.1: Summary of Reliability and SEM Estimates for the Listening and GCVR Sections

	Listening Section		GCVR Section	
Year	Average Reliability	Average SEM	Average Reliability	Average SEM
2011	0.86	4.42	0.93	3.45
2012	0.85	4.71	0.92	4.12
2013	0.89	4.55	0.94	3.93
2014	0.88	4.53	0.94	3.82

estimates for both sections are consistently above the acceptable value of 0.80. Additionally, the SEM estimates as a proportion of the 100-point scale are very small. Overall, these reliability and SEM values suggest excellent consistency of measurement for the MELAB listening and GCVR sections.

In the case of performance tests such as the writing and speaking sections of the MELAB, 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 performance for these sections. The examiners for the speaking test are native or highly proficient nonnative speakers of English who are trained and certified according to standards set by CaMLA. Because the MELAB speaking test is administered by only one examiner, it is not possible to obtain rater agreement figures. Instead, recordings of speaking tests are sent to CaMLA for review.

The raters for the writing section are native speakers of English who are 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 or adjacent agreement on the score to be awarded, the writing performance is evaluated separately by a third rater. The final score awarded for each MELAB essay is the result of exact or adjacent 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 performance.

CaMLA monitors rater agreement for training purposes. The percentage of exact and adjacent agreement between the first and second raters is monitored, along with the Pearson correlation coefficient (r). Table 6.3.2 lists the average rater agreement figures for each year. It shows that the average % Exact + Adjacent agreement between raters is consistently above 0.80, and is often above 0.90. The average Pearson correlation coefficient is also consistently near or above 0.80. Both of these rater agreement figures are high, which suggests excellent agreement among raters. It should be noted that CaMLA also reviews a percentage of essays each year as part of its ongoing quality control processes.

Table 6.3.2: Summary of Rater Agreement Figures for the Writing Section

Year	Average Exact + Adjacent Agreement	r
2011	89.67	0.79
2012	95.90	0.81
2013	88.26	0.82
2014	91.82	0.85

6.4 Trends in Subtest Correlations

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

The MELAB captures evidence of a test taker's communicative language ability through a variety of tasks in the four language skills of listening, reading, speaking, and writing. Section 3.4 described the skills and abilities expected for each language skill, as well

as for grammar and vocabulary knowledge. Test takers are issued a score report that presents their results for each test section. Reporting scores in this way is justifiable if the sections have some overlap (i.e., that they all measure language proficiency) and if each section can also be seen to contribute differentially to the overall MELAB result. Table 6.4 presents the subtest correlations (Spearman's rho) for each year.

Table 6.4: Subtest Correlations (ρ)[1]

Sections	2011	2012	2013	2014
Listening/GCVR	0.70	0.72	0.82	0.83
Listening/Writing	0.54	0.54	0.67	0.67
Listening/Speaking	0.48	0.53	0.55	0.46
GCVR/Writing	0.71	0.67	0.77	0.75
GCVR/Speaking	0.47	0.48	0.56	0.48
Writing/Speaking	0.46	0.40	0.51	0.46

The correlations range between 0.40 and 0.83, indicating a moderate to strong relationship between the subtests. Since each subtest measures language proficiency from a different perspective, these numbers are unsurprising. The moderately strong correlation between the GCVR and writing sections is credible since they both measure use of English either in the form of morpho-syntactic descriptors in the rating scale (writing section) or by explicitly testing grammar and vocabulary (GCVR section). The consistently high correlation between the listening and GCVR sections is more surprising, however, and suggests substantial overlap in the construct elements being assessed. This merits further investigation.

7. Additional MELAB Validity Evidence

Sections 2.2 and 3.4 presented a proposed interpretation of the MELAB score report. 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 MELAB along with the research evidence available for these claims.

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

Proposed claim	Research evidence available
The structure of the MELAB is consistent with its stated construct and with the way in which scores are reported.	• Aryadoust, V. and Goh, C. C. M. (2013) Exploring the relative merits of cognitive diagnostic models and confirmatory factor analysis for assessing listening comprehension, in Galaczi, E. and Weir, C. J. (Eds.) *Exploring language frameworks: proceedings of the ALTE Kraków Conference, July 2011*, Cambridge: Cambridge University Press, pp. 405–426. • Goh, C. and Aryadoust, V.S. (2010). *Investigating the construct validity of the MELAB listening test through the Rasch Analysis and Correlated Uniqueness Modeling*, Spaan Working Papers. • Eom, M. (2008). *Underlying factors of MELAB listening constructs*, Spaan Working Papers. • Wang, S. (2006). *Validation and invariance of factor structure of the ECPE and MELAB across gender*, Spaan Working Papers.

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

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

The language knowledge, processes, and strategies that test takers use to complete MELAB tasks are consistent with the language knowledge, processes, and strategies used by students in college and university settings.	• Gao, L. (2006). *Toward a cognitive processing model of MELAB reading test item performance*, Spaan Working Papers. • Gao, L. and Rogers, W.T. (2011). Use of tree-based regression in the analyses of L2 reading test items, *Language Testing*, 28(1): 77–104. • Li, H. (2011). *A cognitive diagnostic analysis of the MELAB listening test*, Spaan Working Papers. • Song, X. (2005). *Language learner strategy use and English proficiency on the Michigan English Language Assessment Battery*, Spaan Working Papers.
The rating scales for the speaking and writing sections of the MELAB appropriately distinguish between test takers with different levels of language proficiency	• Johnson, J. and Lim, G. (2009). The influence of rater language background on writing performance assessment, *Language Testing*, *26*(4): 485–505. • Jung, Y. J., Crossley, S. A., and McNamara, D. S. (2014). *Linguistic features in MELAB writing task performances*, Poster presented at the meeting of the American Association for Applied Linguistics (AAAL), Portland, Oregon, March 2014. • Lim, G. (2011). The development and maintenance of rating quality in performance writing assessment: a longitudinal study of new and experienced raters, *Language Testing*, *28*(4): 543–560.
The MELAB provides test takers with equal opportunities to demonstrate their language proficiency.	• Aryadoust, V., Goh, C. C. M., and Lee, O. K. (2011) An investigation of differential item functioning in the MELAB listening test, *Language Assessment Quarterly*, *8*: 361–385. • Chapman, M. (2012). *The challenges of ensuring task equivalence in writing tests: A new approach*. Proceedings of the Korea English Language Testing Association: KELTA 2012 Annual International Conference. Seoul, Korea. • Hamp-Lyons, L. and Davies, A. (2006) *Bias revisited*, Spaan Working Papers. • Jiao, H. (2004). *Evaluating the dimensionality of the MELAB*, Spaan Working Papers. • Lim, G. S. (2009). *Prompt and rater effects in second language writing performance assessment*. Unpublished PhD dissertation, University of Michigan. • Lim, G. (2010). *Investigating prompt effects in writing performance assessment*, Spaan Working Papers. • Lin, C-K. (2014). *Treating either ratings or raters as a random facet in performance-based language assessments: does it matter?* CaMLA Working Papers, 2014–01. • Park, T. (2006). *Detecting DIF across different language and gender groups in the MELAB essay test using the logistic regression method*, Spaan Working Papers. • Spaan, M. (1993). The effect of prompt in essay examinations, in D. Douglas and C. Chapelle (Eds.) *A new decade of language testing research: selected papers from the 1990 Language Testing Research Colloquium*, Alexandria, VA: TESOL, Inc., pp. 98–122. • Song, X. (2010). Chinese test-takers' performance and characteristics on the Michigan English Language Assessment Battery, in Cheng, L. and Curtis, A. (Eds.) *English language assessment and the Chinese learner*, New York and London: Routledge, Ch. 9. • Yu, G. (2009). Lexical Diversity in Writing and Speaking Task Performances, *Applied Linguistics*, *31*(2): 236–259

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

Performance on the MELAB is related to other indicators of language proficiency in academic contexts.	• Dobson, B., Han, I., and Yamashiro, A. D. (2001). *The relationship between test scores on the MELAB and the TOEFL CBT*, UMELIRR2001-01, Ann Arbor, MI: University of Michigan.
The MELAB 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.

7.1 The structure of the MELAB 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 Examination for the Certificate of Proficiency in English (ECPE) and the MELAB. The MELAB data comprised 216 test takers, all of whom took the same listening and reading (GCVR) forms as well as a writing section of the exam. 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 68% of the test takers in this dataset were female, a finding that is in line with the trends described in Section 5.2. The analyses also showed that the female test takers had a slightly higher mean test score than the male test takers, in line with the findings reported in Section 6.2.

Exploratory factor analyses (EFA) were performed that took into account the different item types represented in the exam. These analyses revealed one dominant factor with an eigenvalue of 3.54 which accounted for 59.1% of the common variance. This result suggested that the MELAB 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 MELAB was 5.69, confirming a single meaningful factor to explain the MELAB data.

After determining that a one-factor model best explained the data, a confirmatory factor analysis (CFA) was performed. This 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 MELAB (when all the sections are taken together in the calculation of a test taker's final result) measures English language proficiency. This supports claims that the MELAB test sections together measure a test taker's overall English language proficiency as well as CaMLA's practice of reporting a MELAB total score that is an average of test takers' performance on all written sections of the examination.

Two subsequent studies have focused on the MELAB listening section. Eom (2008) hypothesized that the MELAB listening section consists of two factors: language knowledge and listening comprehension. The dataset comprised 2,133 test takers, all of whom took the same test version. As a first step in her study, Eom (2008) analyzed the test content and identified 14 listening abilities (10 associated with language knowledge and 4 associated with listening comprehension). She then tested this tentative model using Confirmatory Factor Analysis (CFA). Eom's *a priori* model met most of the fit criteria, the exception being the Chi-square (χ^2) test. This was a satisfactory result and indicated that the MELAB listening section does test language knowledge and listening comprehension. Nevertheless, Eom decided to respecify the model, allowing error terms to covary in order to reduce the standard residuals in the initial model (2008: 89). She argued that listening constructs are complex and that it was possible that factors other than the two she had specified (language knowledge and listening comprehension) exerted an influence upon the model. The respecified model showed better fit,

thus confirming Eom's hypothesis that the listening items test common factors other than the two that she modeled.

Eom's decision to respecify the model by allowing the error terms to covary was later questioned by Goh and Aryadoust (2010) because the error terms were covaried without a theoretical principle. In their study, Goh and Aryadoust (2010) posited three underlying competency-based models for the data. One model contained the following components: minimal context questions, detailed (or explicit) questions, close paraphrase questions, propositional inference questions, and enabling inference questions. The second model contained the same components but allowed the error terms to correlate. The third model posited a two-level latent trait with a higher level listening trait above the components identified in model 1. The dataset comprised 916 test takers, all of whom took the same test version in the period February–August 2008. The number of male and female test takers was approximately equal and they came from a large variety of first language backgrounds (78 different L1s).

Three different analyses were conducted, beginning with descriptive statistics and reliability analyses. The former indicated that the items were normally distributed. The reliability analyses were performed for subgroups of items (identified by the underlying models that had been posited). Each of the subgroups contained a small number of items and this affected the reliability indices achieved. However, the reliability index for the listening section as a whole was a satisfactory 0.85. Confirmatory factor analyses (CFA) were then conducted to examine the fit of the three postulated models. None of these initial models proved a good fit to the data. Goh and Aryadoust (2010) therefore revised their hypotheses. They set aside the competency-based models and proposed an alternative task-based model that corresponded to the three subsections in the listening test: understanding and responding to (i) short, minimally contextualized statements, queries, and requests; (ii) short conversations; and, (iii) extended radio interviews. CFA showed that this model fit the data satisfactorily "testif[ying] to the presence of a firm three-factor construct which underpins" the MELAB listening section (Goh and Aryadoust, 2010: 50).

Next, Goh and Aryadoust (2010) performed a Rasch analysis of the data to explore the item and person fit indices and the dimensionality of the test. A principal component analysis of residuals provided evidence for the unidimensionality of the test and also indicated local independence of the items. Inspection of the item and person fit indices revealed a mismatch between the distribution of item difficulties and the ability of the test takers. The person mean was 0.68 logits above the anchored item mean, and approximately 24% of the test takers did not receive sufficiently challenging items. As expected, the standard error of measurement indices for these test takers is larger because they received fewer items that corresponded to their ability. This finding is partly in accordance with the aims of the test, which is to provide a reliable measure of listening proficiency for test takers who are minimally ready to enter a program of study or to work in an English-medium context.

Finally, Goh and Aryadoust (2010) performed a DIF analysis to examine gender bias. This analysis flagged seven items for DIF, five that were more difficult for female test takers, and two that were more difficult for male test takers. The effect of these items on test takers' listening scores is quite small because they only account for a small percentage of the total number of items. In summary, Goh and Aryadoust (2010) conclude that their analyses provide good evidence for the construct validity of the MELAB listening test. Nonetheless, they suggest that the minimal context items harken back to "an older generation of listening items" (2010: 59). They recommend further investigations to establish whether candidates who get these items incorrect do so because they are unable to understand the meaning of the decontextualized statements (a construct relevant factor) or because there is insufficient context provided (a construct irrelevant factor).

In a subsequent investigation of the underlying structure of the MELAB listening test using cognitive diagnosis modeling (CDM) and Confirmatory Factor Analysis (CFA), Aryadoust and Goh (2013) excluded the minimal context items from their analyses. They argue that these items have "a less communicative structure, which could affect the modeling" (2013: 410). Aryadoust and Goh subjected the remaining 35 items to a detailed content analysis using expert judgment. They identified four listening subskills which they tested using CFA. They found that the four-factor model fit the data well but that the correlation coefficient between *detailed information* and *close paraphrase* was unacceptably high, suggesting that those two subskills overlap and can be combined. Aryadoust and Goh therefore tested a three-factor model in which

those subskills were combined. This model also fit the data well but revealed another "offending correlation" (2010: 415) which led to Aryadoust and Goh testing a two-factor model. Each of these models fit the data well but each of them contained an overly strong correlation that suggested that the skills could be further combined. Since all the items test listening, this is to be expected. However, these results also show that CFA is unable to explain the underlying structure of the data.

Aryadoust and Goh suggest that this might be because CFA is inappropriate for dichotomously scored tests which, perhaps artificially, group test takers into "masters" or "nonmasters" for each item. They found fusion modeling (a type of CDM) far more promising for the MELAB listening dataset. The model conformed to the underlying traits identified by the item-content analyses and provides support for the validity of the listening section.

7.2 The language knowledge, processes, and strategies that test takers use to complete MELAB tasks are consistent with the language knowledge, processes, and strategies used by students in college and university settings

Several different studies were conducted that examined the language knowledge, processes, and strategies used by test takers to complete the MELAB. Song (2005) conducted a study that examined language learner strategy use and English proficiency on the MELAB. The study participants consisted of 161 MELAB test takers who took the exam and provided valid responses to a strategy use questionnaire. The questionnaire consisted of 43 items (27 on cognitive strategy use and 16 on metacognitive strategy use), each of which used a six point Likert scale (0–never, 1–rarely, 2–sometimes, 3–often, 4–usually, and 5–always). The study aimed to examine the nature of language strategies reported by test takers and to investigate the relationship between the reported strategies and MELAB performance.

Exploratory factor analysis was performed on cognitive and metacognitive strategy use items separately to examine the underlying factors. It found that cognitive strategy use had six underlying factors and metacognitive strategy use had three underlying factors. The cognitive strategy use covered six dimensions: *repeating/confirming information strategies, writing strategies, practicing strategies, generating strategies, applying rules strategies,* and *linking with prior knowledge strategies*. The metacognitive strategy use had three dimensions: *evaluating, monitoring,* and *assessing*. These results were partially consistent with Purpura's (1999, cited in Song, 2005) framework. Song (2005) listed several potential reasons for the differences, such as the small number of participants, how some items were worded, and the fact that this study was conducted in an ESL context.

Regression analysis was performed to examine the relationship between learner strategy use and MELAB scores. Stepwise regression was performed for the strategies and the MELAB listening, GCVR, writing, and total scores. It found that not every type of strategy use had a beneficial effect on MELAB scores. Several strategies (*Applying rules, practicing, assessing,* and *evaluating*) had no significant effect on any section of the MELAB. Others had significant positive or negative impacts on test scores for some sections, but not for others. While the effectiveness of the strategies varied based on the test section, the study found that *linking with prior knowledge* consistently showed a significant positive effect, and that *repeating/confirming information* consistently had a significant negative effect. Overall, the study provided evidence of a linear relationship between strategy use and the MELAB, but the effect was weak. Strategy use only explained 18.9% of the variance in the MELAB total scores, 21.4% of the variance in writing scores, 17.2% of the variance in listening scores, and 12.5% of the variance in GCVR scores. Song (2005) concludes that these results are consistent with those of other studies, and that strategy use should only explain a small proportion of the variance since it is only one piece that affects performance on the MELAB.

The next two studies used cognitive processing models and tree-based regression (TBR) analysis on MELAB reading items. Gao (2006) developed and tested a cognitive processing model that was hypothesized to underlie MELAB reading item performance. Gao and Roger (2011) explored whether the results of TBR analysis, informed by a valid cognitive model, would enhance the interpretation of the cognitive processes involved in answering MELAB reading items based on item difficulty. The two studies are very similar, using the same test data and similar research methods. The study by Gao and Roger (2011) appears to be an attempt to replicate and improve upon the work done in Gao (2006).

Both studies utilized the same test forms and test taker performance data. The studies used test taker responses to the reading sections of two parallel MELAB forms (E and F). Each form contained 20 multiple choice items based on four reading passages. The procedure for performing the analysis was also generally the same for both studies. First, an initial cognitive model was hypothesized to underlie the MELAB reading item performance based on a review of the relevant literature. Gao (2006) developed an initial model that consisted of 10 cognitive elements, while Gao and Rogers (2011) developed an initial model that consisted of 14 cognitive elements. Next, the MELAB reading items were reviewed by 3 trained raters with experience in L2 reading who coded the items based on the cognitive processes required to correctly answer each item. Gao (2006) had 3 doctoral students in educational psychology code the items, while Gao and Rogers (2011) had two psychometricians and an applied linguist code the items. The MELAB reading items were then administered to a number of Chinese/Mandarin L1 students (10 in Gao, 2006; 16 in Gao and Rogers, 2011) who reported verbally what they thought while they were answering the items. These verbal reports were analyzed and coded for the cognitive processes that the students used to answer each item. The results of both coding steps (item ratings and verbal reports) were closely examined, and were used to modify and validate the initial cognitive models. The validated cognitive processing models were then analyzed using TBR analysis for each test form to see to what extent the cognitive processes used to correctly answer MELAB reading items explained item difficulty.

TBR analysis is a regression technique that outputs a regression tree that can be used to identify the cognitive processes required to solve each item in a test. The two studies differed somewhat in this piece of the analysis. In Gao (2006), the validated cognitive model consisted of nine elements; however, not all of them were used as predictors in the TBR analysis for each form. The Form E tree had four predictors (*evaluating alternative options, drawing inferences, using pragmatic knowledge,* and *processing academic text with specialized and infrequent words*) that explained 90.7% of the total variance in item difficulty, and the Form F tree had three predictors (*evaluating alternative options, drawing inferences,* and *using syntax knowledge*) that explained 94.5% of the total variance in item difficulty. In Gao and Rogers (2011), the validated cognitive model consisted of ten elements, all of which (*word recognition, vocabulary knowledge, syntactic knowledge, knowledge of discourse structure, synthesis, drawing inferences, pragmatic knowledge, locating specific details in text, matching questions to text, and evaluating alternative choices*) were utilized as predictors in the TBR analysis for each form. The Form E tree explained 97.9% of the total variance, and the Form F tree explained 99.3% of the variance. It should be noted that Gao and Rogers (2011) did not apply any pruning to the trees since the purpose of the analysis was to determine what reading and test management processes affect item difficulty. They also state that while the final trees may be over fitted, they provide an indication of the processes measured by each item in each form.

In both studies, the results of the TBR differed between forms, but the similarities in the final trees shed light onto which construct-relevant item features most likely affected the MELAB results. The results of both Gao (2006) and Gao and Rogers (2011) suggest that items with more plausible distractors are more difficult than items with fewer plausible distractors. Furthermore, the studies revealed that the number of plausible distractors was the strongest predictor, which was in line with the findings from the literature. Both studies also suggest that items that require high-level inference tend to be more difficult than items that do not require it. Overall, both studies were successful in using cognitive processing models and TBR analysis to examine MELAB reading items. Gao (2006) was able to conclude that the cognitive process model developed and tested in the study linked the theory in the domain of L2 reading to the MELAB reading items, and Gao and Rogers (2011) were able to conclude that the study reveals that TBR can be used to enhance the interpretation of a statistical item analysis in terms of the cognitive processes used by students when responding to L2 reading items. Both of these studies work to show that the reading processes and strategies used by test takers to complete the MELAB are consistent with those found in the theory.

In another study, Li (2011) performed a cognitive diagnostic analysis of the MELAB reading section. Cognitive diagnostic models (CDMs) provide test takers with multidimensional skill profiles that classify them based on whether or not they have mastered each of the skills involved in the exam. Specifically, this study used the Fusion Model to estimate the test taker profiles on the reading subskills underlying the MELAB reading section. Like other CDMs, the Fusion Model is a confirmatory and multidimensional latent-variable

model, however, it differs in that it contains a residual parameter that helps compensate for incomplete knowledge about the skills required to answer items. Li (2011) argues that cognitive diagnostic analysis is useful because it provides more detailed information than traditional IRT analysis, and the diagnostic feedback can be used to facilitate better teaching and learning. The purpose of the study was to investigate the use of cognitive diagnostic analysis with the MELAB reading section so that more detailed diagnostic information could be provided to MELAB test takers.

This study used the item response data of 2,109 test takers to the reading section of MELAB Form E. The form contained 20 reading items based on four passages. The first step in the analysis was the development of the initial loading structure of the Fusion Model, known as the Q-matrix. This matrix contains the skills that are hypothesized as required to successfully answer each item. This study's initial framework included the skill categories: *vocabulary, syntax, extracting explicit information, connecting and synthesizing*, and *making inferences*. These categories were selected based on the classifications of reading subskills found in other studies, including Gao's (2006) study of the MELAB reading section. This initial framework was examined further through analysis of data from think-aloud protocols and expert rater. For the think-aloud protocols, thirteen ESL learners were administered the test and reported verbally what they were thinking for each item. The transcribed recordings were then closely examined to determine what reading skills were involved. For the expert rater analysis, four experts with experience in teaching reading to ESL students examined each of the 20 reading items to identify the skills required to provide the correct answer. These analyses mostly supported the initial framework, so the framework of the initial Q-matrix was not modified.

The item response data for each of the 20 reading items was used to empirically validate the Q-matrix by using the Markov Chain Monte Carlo (MCMC) algorithm to check for convergence. It found that the overall convergence of the initial Q-matrix was acceptable, but could be improved. Only three items required test takers to *make inferences*, which is too small of a number for sufficient estimation. Therefore, a revised Q-matrix was analyzed that condensed *making inferences* and *connecting and synthesizing skills* into one skill: *understanding implicit information through connecting ideas and making inferences*. This refined matrix converged better than the initial one. The model fit of the initial and refined Q-matrices were compared, and it was found that both matrices resulted in a reasonable fit. Because the refined Q-matrix was more parsimonious, it was selected for use in the Fusion Model. This model was then used to estimate the item parameters and examinee skill mastery status. The model showed that 25.6% of the test takers mastered *vocabulary*, 28.7% mastered *syntax*, 40.1% mastered *extracting explicit information*, and 32.3% mastered *understanding implicit information through connecting ideas and making inferences*. Examining the skill profiles together revealed that 16.05% of the test takers mastered all four skills, and 51.61% were nonmaster for all four skills. Li (2011) concludes that this study was able to provide useful diagnostic information for MELAB test takers. However, Li (2011) also cautions that cognitive diagnostic modeling is new to the field of language assessment, so further investigation of the technique is required. It is also noted that applying CDMs to existing exams is not an optimal approach. Detailed diagnostic analysis would work best if the exam was initially built for this kind of skill based assessment purpose.

7.3 The rating scales for the speaking and writing sections of the MELAB appropriately distinguish between test takers with different levels of language proficiency

Three different research studies have been done that examine the appropriateness of the rating scale for the MELAB writing section. The first study, by Johnson and Lim (2009), investigated the influence of rater language background on the writing assessment for the MELAB writing section. The study examined essays from 7,400 MELAB test takers who represented 21 different first language backgrounds. Because MELAB essays are scored at least two times, there were a total of 15,635 ratings to examine. The essays examined were scored by 17 different raters, 13 of which were native speakers (NS) of English, and 4 of which were nonnative speakers (NNS) of English. Two of the NNS raters were native Spanish speakers, one was a native Korean speaker, and another was a native Amoy and Tagalog speaker. Due to the small number of NNS examined in this study, Johnson and Lim (2009) caution against generalizing the findings of this study to other NNS raters. The purpose of this study

was to examine whether the language background of a rater had any effect on the rating of MELAB writing essays, and if so, to determine if there was any pattern of interaction between the rater and test taker first languages.

The IRT program, FACETS, was used to analyze this data for rater severity, fit, and language background bias. Overall, the analysis showed that there was no pattern of bias in the ratings related to rater language background, and that NNS raters were not found to be any more or less harsh than NS raters. While the analysis contained a few significant bias terms, Johnson and Lim (2009) note that the magnitudes of these terms were generally too small to be substantial. Based on this analysis, the authors were able to say that there was sufficient evidence to conclude that the rater language background did not affect the scoring of the MELAB writing section.

In another study, Lim (2011) investigated the performance of new and experienced raters on the MELAB writing section over time. The data for this study covers three distinct time periods and consisted of 20,662 ratings from 11 raters. Of these raters, 5 were experienced raters that were in every time period. Two new raters were introduced each time period, which means the study contained a total of six new raters. It is important to note that over time, the new raters became reclassified as experienced raters. This means that there were never more than two new raters per time period. The purpose of this study was to examine how the rating quality of novice MELAB raters develops over time, and to see to what extent all MELAB raters maintain their rating quality over time.

FACETS was used to perform multi-facet Rasch analysis of the data. The analysis was done monthly for each time period so that rater performance could be observed over time. It specifically looked at the measures of rater severity and consistency. Overall, the analysis showed that novice raters often performed no differently than experienced raters, and that in cases where the novice raters did perform worse, the improvement in their rating quality occurred relatively quickly. The report notes that while the reasons for this improvement could not be determined from the study, the results suggest that the frequency or number of ratings done may be an important factor that influences the quality of the ratings. The study found that all of the novice raters had acceptable rater performance by the 130th essay rating. Lim (2011) proposes that this finding supports the validity of the MELAB policy that raters score at least 80 essays before they can become fully certified. The study also showed that raters were able to maintain a consistent level of quality over time, which supports the idea of an expert rater. However, Lim (2011) also cautions that the raters included in this study rated regularly, and that even then, one rater briefly became inconsistent over a period of time when she evaluated a small number of essays. This suggests that there may be a minimum amount of continuous experience required to maintain a rater's quality.

In a third study, Jung, Crossley, and McNamara (2014) examined the linguistic features elicited by the MELAB writing tasks. This was done using the program Coh-Metrix to identify the linguistic features, and then by performing a regression to determine which ones were significant predictors of MELAB essay scores. The researchers had a corpus collection of 1,003 MELAB essays that were administered in 2013. The sample was stratified according to score level, gender, and age so that is was representative of the MELAB test population. The data analyzed for this poster presentation was a sample of 500 essays from this corpus. It was divided into a training set (n = 334) and a test set (n = 166) for the analysis. The purpose of this study was to determine what linguistic features distinguished MELAB test taker writing performances on the MELAB rating scale.

Coh-Metrix was used to analyze the essays in the training set for several linguistic features. In total, 18 indices were selected for use in the regression analysis. Of these 18 indices, 9 were found to be significant predictors of the essay scores. They were: *number of words per text, word frequency (content words), lexical diversity, word meaningfulness, semantic similarity (latent semantic analysis paragraph to paragraph), number of modifiers per noun phase, content word overlap, number of words before main verb*, and *causal connectives*. This model explained 55.1% of the variance in the rater scores for the training dataset. The regression model was also able to be extended to the test dataset, where it explained 54% of the variance in rater scores. This analysis provided evidence that linguistic features can predict ratings on MELAB essays. The linguistic features with the greatest predictive value were associated with text length and lexical sophistication. Jung, Crossley, and McNamara (2014) conclude that this analysis contributes to the validity of the raters use of the MELAB rating scale by verifying the linguistic features used in the scale.

7.4 The MELAB provides test takers with equal opportunities to demonstrate their language proficiency

A variety of studies have investigated the fairness and equity of the MELAB for test takers. Some of these studies have compared the performance of a subgroup of test takers with the population as a whole. Others have explored differential item functioning for one or more sections of the exam. A third type of study has investigated the equivalence of different forms of the exam. Finally, some studies have evaluated the effects of rater training for the writing section.

Song (2010) describes the performance of 174 Chinese L1 test takers who took the MELAB in 2004 and compares their characteristics to the total MELAB population for that year. She reports that, for the speaking test, the Chinese L1 subgroup received the same average score as the total MELAB population. However, for the listening, reading, and writing sections, the subgroup received a slightly higher average score than the total MELAB population. This difference, however, was very small (the mean final score was 0.72 of a point higher) and might not be meaningful. Within this subgroup of Chinese L1 test takers Song (2010) also inspected differences in MELAB part and final scores by gender, purpose for taking the test, age, and country of origin. She found differences between groups for each of these test-taker characteristics but the differences were small and not statistically meaningful. This interesting, albeit small-scale study shows that the MELAB was fair for this subgroup of Chinese L1 test takers and did not unwittingly bias towards or against them.

Building on an earlier study of the internal structure of the MELAB listening test (Goh and Aryadoust, 2010), Aryadoust, Goh, and Lee (2011) investigated DIF in the same dataset, hypothesizing that some items in the test would demonstrate gender-based DIF. As a preliminary step in this study Aryadoust et al. (2011) calculated the descriptive statistics for the dataset, performed a Rasch analysis and reliability analysis, and tested the data for unidimensionality and local independence. These preliminary analyses confirmed that the MELAB listening dataset is very reliable and also satisfied the preconditions for DIF analysis. The subsequent uniform DIF analysis (UDIF) identified eight test items with significant DIF. Of these, two items demonstrated nonuniform DIF (NUDIF); the items favored different subgroups (male or female test takers) depending on their ability levels. To explore this finding in more detail, Aryadoust et al. (2011) divided the gender subgroups into high- and low-ability classes. They then performed a NUDIF analysis of all the test items. This revealed 22 instances of significant NUDIF (2011: table 4) for 15 items. Further analysis of these items showed that, in some cases, the DIF could be attributed to construct-relevant factors. For instance, items 30 and 35 tested the ability of test takers to identify an accurate paraphrase of the stimulus. Wagner (2004) has shown that high-ability test takes perform better in this skill than low-ability learners. These items demonstrate and confirm Wagner's findings. In the case of the other items, Aryadoust et al. (2011) generated the following explanatory hypotheses:

- Low-ability male test takers might use guessing as a strategy for answering difficult questions. Items where this strategy is successful demonstrate DIF in favor of this subclass of test takers.
- If one or more of the distracters for a difficult item is very attractive, this dissuades the low-ability male test takers from making guesses. The resulting high discrimination figures may be the cause of DIF in favor of the high-ability test takers.

In other words, the analyses show that in many of the items where DIF was identified (10 items), the DIF was caused by a confounding variable external to the test—test-taking strategies used by a particular subclass of test takers. This is construct-irrelevant variance; Aryadoust et al. (2011) suggest this "noise" could be minimized by improving the distracters in order to make them more attractive to test takers. This improvement to item-writing procedures has already been implemented. It would therefore be useful to repeat this study in order to confirm that this area of construct-irrelevant DIF has, indeed, been reduced or eliminated.

The MELAB offers test takers a choice of two writing prompts. This approach allows test takers to answer on a topic that is of most interest to them and is sensible if the test takers' final writing test score is not affected by prompt choice. Spaan (1989, cited in Spaan, 1993) investigated writers' prompt choice. Study participants were assigned one of two MELAB prompt sets. Each prompt set contained a challenging prompt; that is, a prompt that had "more sophisticated content and more rhetorical specification" (Spaan, 1993: 101). The alternate prompt in each prompt set was predicted to be less challenging. Spaan (1989) found that in both

groups the less challenging prompt was more popular. However, when Spaan investigated the writing scores for each prompt she found that for one prompt set the writers who chose the less challenging prompt scored significantly better whereas for the second prompt set, the writers who chose the more challenging prompt scored significantly higher. Using the same MELAB prompt sets, Spaan (1993) explored these findings in more detail. Her research questions were as follows:

- Does performance differ when prompts differ?
- If so, how can the differences be measured—through holistic scores, test analysis, or both?
- If so, do the differences occur at different ESL proficiency levels, difference academic levels, or both?
- If performance differs in writing tasks in which students may select a topic, do they choose their optimal topic (the prompt on which they write better)?

The prompts were analyzed using a scheme developed by Vahapassi (1982) and Purves, Soter, Takala, and Vahapassi (1984) and then categorized as either Narrative/Personal (NP) or Argumentative/Impersonal (AI). The dataset comprised 88 test takers, each of whom wrote on two prompts (one NP and one AI) and also took the listening and reading sections of the MELAB. Prompt order was varied to offset practice effects. The test takers also completed a post-test questionnaire. Each essay was evaluated by a minimum of two examiners.

Spaan (1993) first established that the study group was comparable in language proficiency to a typical MELAB population. She also established that subgroups within the population performed as expected (for instance, the graduate student group performed better on the exam than the undergraduate student group). These analyses established the representativeness of the study group. Spaan (1993) then compared the test takers' performance on the two essay prompts. In the majority of cases (79, 90%), the test takers scored the same on both prompts; any differences were small and attributable to chance. The nine test takers whose essay scores differed by more than one ranking (labeled "inconsistent" scorers) were matched with 12 "consistent" scorers according to proficiency level, part scores, language, and native country. This resulted in a data subset of 42 essays (21 test takers, 2 essays per test taker) on which detailed analyses were performed for fluency, syntactic sophistication and accuracy, lexical range and sophistication. Though cautioning overinterpretation of the results because of the small sample, Spaan (1993) found that the test takers' writing was strikingly similar regardless of the prompt they answered. She also found that the AI prompts resulted in shorter essays than the NP prompts but that, overall, holistic scores were not affected by prompt answered. Spaan (1993) concluded that offering a choice of prompt does not appear to be detrimental to test takers but she pointed out that one of the AI prompts demanded far more specific content knowledge. This prompt generated the shortest, least developed essays, leading Spaan (1993) to recommend that subject content should be more accessible (or universal) in the future.

Spaan's (1993) work provides good evidence that the MELAB practice of giving test takers a choice of two writing prompts is fair. Test takers' writing scores are not affected (positively or adversely) by the type of essay prompt they write on—be it a narrative, argumentative, or expository prompt. This study also indicates that the linguistic features of a test taker's writing are similar regardless of the prompt selected, suggesting in turn that the inferences made about the test taker's writing ability can be the same regardless of prompt selected.

Park (2006) and Lim (2010) also explored the comparability of MELAB essay prompts, looking at the comparability of prompts across test administrations.

Park (2006) was interested in whether MELAB writing prompts allow test takers to show their ability regardless of their language background[2] or gender. The data comprised 2269 test takers who had each answered one of 10 different writing prompts (N per prompt > 140). The test takers' results for the MELAB listening and reading sections were used as an independent measure of their English language ability. Park first checked whether the different language and gender groups were equal in ability (as measured by the MELAB essay score and the English language ability score). He detected small standardized mean differences in both the essay and the English language ability scores for the two language groups. However, the two gender groups were indistinguishable in ability. Park then used independent samples t-tests to explore

2 The languages were grouped into Indo-European languages (including Hinidi, Punjabi, French, Portuguese, and German) and non-Indo-European languages (including Tagalog, Japanese, Tamil, and Chinese).

group differences by prompt. This analysis revealed significant differences by language group for eight of the ten prompts and significant differences by gender for three of the ten prompts. Park then used logistic regression to evaluate the practical significance of these differences. He found that the R^2 effect sizes were too small for any prompt to be identified as demonstrating DIF for language or gender. It appeared that the score differences were due to "item impact" (2006: 92); the test takers had different probabilities of performing well on the prompts because they differed in the underlying ability measured by the prompt. This finding supports the validity of MELAB writing score interpretations.

In an unpublished PhD dissertation, Lim (2009) investigated the effects of different writing prompts and different raters on MELAB writing scores. Part of this dissertation was later published in Lim (2010), which investigated only the effects of different writing prompts on MELAB writing scores. These two studies utilized the same dataset and research method, and arrived at the same conclusions. The dataset included 29,831 ratings of 10,536 test takers on 60 different prompts. The ratings were analyzed by a total of 24 different raters. FACETS was used to perform multi-facet Rasch analysis of the data. Lim (2009) posed six research questions, the first three of which are in Lim (2010). They were:

- Are the prompts comparable in difficulty, and does the test have a prompt effect?
- Is there a prompt effect relating to topic domain, rhetorical task, prompt length, task constraint, expected grammatical person of response, and number of tasks?
- Is there a bias effect between the writing prompts and test taker gender, language background, or proficiency level?
- Do the raters rate appropriately and consistently, and does the test have a rater effect?
- Is there a rater effect relating to experience, time, or language background?
- Is there a bias effect between the raters and the prompt difficulty or the prompt selection?

With regard to test prompts, the study found that the prompts were comparable in difficulty, and that any differences in prompt difficulty were generally not large enough to affect scores. It also found that there was no interaction effect between the writing prompts and the test takers' gender, language background, or proficiency level. Of the prompt and test taker dimensions examined, only prompts on social topics appeared to be difficult enough to make a statistically significant difference in MELAB writing scores. However, the effect is small in magnitude, accounting for less than 0.15 of a scale point. Regarding the raters, the study found that the MELAB raters were trained to rate appropriately and consistently, and that rater language background had no effect on scores. While newer raters were somewhat more variable in their scores, the process of becoming an experienced rater occurred in a short amount of time. Once raters gained experience they had relatively stable leniency/severity and that they were consistent in their ratings. The study also found that there was no bias effect between raters based on prompt difficulty or selection. Overall, the results of Lim (2009) provide evidence that assigning different prompts to different test takers and assigning different raters to different test takers did not affect the validity of the scores, and that the MELAB results are still valid, reliable, and fair

Chapman (2012) explored the equivalency of MELAB writing tasks through qualitative analysis of MELAB prompts and essays. He worked to determine the distinguishing characteristics of the writing prompts, and the effects of these characteristics on the both the writing process and the essay. The study was conducted in three stages in order to (1) identify the distinguishing characteristics, (2) gain insight into the test takers' thought process when selecting a prompt, and (3) investigate how the characteristics of the prompt affected the linguistic qualities of the essay. This was done through analysis of the prompts and essays, and through test taker interviews.

The first stage of the study identified a total of four distinguishing characteristics of MELAB writing prompts: domain (educational, occupational, public, or personal), response mode (narrative or argumentative), number of rhetorical cues (defined as instruction in prompt that writer must respond to), and open or focused (do test takers require little to no background knowledge, or does the prompt need contextualization). The second stage of the study revealed that test takers were influenced by the exam's time constraints when selecting the prompt to write about. The test takers tended to gravitate towards personal domain questions since the topics were more familiar and therefore perceived as easier. The response mode and the open or focused nature of the prompts did not appear to have any effect on the decision making process, although

the study notes that this may be because of the close relationship between these characteristics and personal domain prompts. Prompts with more rhetorical cues, however, were found to be more helpful to the test takers when composing the essay because they provided them with guidance on how to structure the content of the essay. The third stage of the study was primarily exploratory in nature, but it found that the personal domain prompts seemed to elicit linguistically different responses from the other domain prompts. Essays on personal domain prompts tended to have lower frequency lexis that had more richness and better cohesion at lower language proficiency levels than other prompt domains.

Overall, Chapman (2012) suggests that because of the different language elicited by personal domain prompts, they should only be used for high-stakes testing if they are specifically the construct that is to be measured. He concludes that while it is important for writing tasks to sample from a broad range of topic domains to avoid becoming too predictable, it is also important that test developers have a clear construct definition of the writing proficiency that is to be measured.

Jiao (2004) investigated the dimensionality of the MELAB listening and GCVR sections separately using item level information, and also examined the influences of gender (male or female), native language (Korean or Tagalog), and proficiency level (high or low) on the dimensionality. This study utilized two sets of test taker response data. The first dataset was from MELAB Form FF, and contained the response data for 1,650 test takers on the listening section. The second was from MELAB Form EE, and contained response data for 1,031 test takers on the GCVR section. The dimensionality of each section was assessed using two different techniques: Stout's nonparametric analysis of dimensionality, and principle component analysis with tetrachoric correlations. Stout's nonparametric procedure tests the null hypothesis of one dimension against the alternative of multiple dimensions. Principal component analysis is used to assess the dimensionality of a set of items, using multiple criteria (eigenvalue, proportion of variance accounted for, and interpretability criterion) to determine the number of components (dimensions) to retain. The author notes that because Stout's procedure is more effect for larger sample sizes ($n \geq 2{,}000$), caution should be taken when generalizing the results.

The study found that the dimensionality of the MELAB was somewhat inconsistent between the two sections and between the different subgroups. While the GCVR section was found to be unidimensional for the general sample population, the listening section was not. When the dimensionality of the different subgroups was examined, the study found that both the listening and GCVR sections were unidimensional for only female and Tagalog speaking test takers. Subgroups of test takers who were male, Korean speaking, high proficiency, or low proficiency were not unidimensional. Overall, Jiao (2004) concludes that the results of this study help to identify the effect of MELAB items and test taker characteristics on the dimensionality of the MELAB listening and GCVR sections. Given these results, it would be interesting to perform these analyses for other MELAB administrations to better understand the dimensionality of the listening and GCVR sections.

Hamp-Lyons and Davies (2006) conducted an investigation on the MELAB writing section of bias in relation to the International English (IE) view and the World English (WE) view. The IE view holds that the English of an educated native speaker is the norm to which all others should be compared. The WE view holds that local standards are already in place, and that imposing the IE view on users of WEs is potentially discriminatory against nonnative English speakers. A total of 60 essays were used in this study; 10 from each of the following language backgrounds: Arabic, Bahasa, Japanese, Chinese, Tamil, and Yoruba. In addition to original MELAB ratings, each essay was rated using a simplified TOEFL writing scale by pairs of raters from each of the six language backgrounds. The purpose of this study was to compare the MELAB writing scores from raters and test takers with a shared language background with their scores from certified MELAB raters. The authors hypothesized that if there was bias, it would be reflected in a significant difference between the scores of native speaking raters' and MELAB raters.

While the premise of this study was rather interesting, the results were inconclusive, primarily because there were too few test takers for each language background. Inconsistencies with the raters used in the study, and the use of two different rating scales to score the MELAB essays also could have affected the results. Overall, while this study does not provide evidence to support the validity of the MELAB, it does provide a framework for a study that would be interesting to replicate with a larger sample size.

Yu (2009)[3] examined the relationship between lexical diversity and test taker performance on MELAB writing and speaking tasks using *D* as the measure of lexical diversity. The study utilized a total of 200 writing compositions and 25 speaking interviews collected from MELAB exams administered between January 2004 and November 2005. The 25 speaking interviews were selected from the 200 candidates whose writing compositions were already selected. Care was taken to ensure that the sample contained a good range of scores, and that the writing compositions represented only five prompts: two of which were personal and three of which were impersonal. In this study, Yu (2009) examined five main areas:

- The relationship between the lexical diversity of written discourse and the writing score.
- The effects of different writing topics on the lexical diversity of written discourse.
- The relationship between the lexical diversity of spoken discourse and the speaking score.
- The relationship between the lexical diversity of the written and spoken discourses and test taker overall proficiency.
- The relationship between the lexical diversity of written and spoken discourses.

After obtaining estimates of lexical diversity (*D*) for each writing and speaking performance, Yu (2009) examined the relationships listed above using linear regression and correlation analysis. The study found that there was a significant positive relationship between lexical diversity and both the writing score (r = 0.33) and the speaking score (r = 0.48). The results show that lexical diversity was able to predict speaking performances better than writing performances, with the lexical diversity explaining 23% of the variance in speaking ratings and 11% of the variance in writing ratings. Yu (2009) notes that these coefficients are quite high considering the numerous other factors that can affect a rater's judgment of writing and speaking performances. The study also reveals that the lexical diversity of the writing and speaking performances are positively correlated (r = 0.45) and have approximately equal *D* values. Analysis of the relationship between lexical diversity and overall test taker proficiency revealed that the final MELAB score had a significant effect on the lexical diversity of writing and speaking, explaining 9.3% of the variance in *D* for writing performances and 24.7% of the variance in *D* for speaking performances. Using the scores on the listening and GCVR sections of the MELAB as measures of test taker overall proficiency instead of the final score produced similar results. Finally, the effects of the five writing topics on lexical diversity were analyzed using ANOVA, which found that there was a statistically significant difference in *D* for the different writing prompts. Upon closer investigation of the topics and demographic characteristics of the test takers, the author concludes that this difference had to do with some groups (i.e., nursing students) being more familiar with some essay topics (i.e., plastic surgery). Overall, Yu (2009) showed that lexical diversity had statistically significant positive correlations with test takers' general language proficiency and performance on the MELAB writing and speaking sections. The consistent significance of lexical diversity in writing and speaking performances provide evidence for the validity of including lexical diversity as a quality indicator in the MELAB rating scales.

Lin (2014) compared two different methods of estimating variance components in performance-based assessments for sparse datasets. The rater method treats the raters as a random facet, and works by first identifying all fully crossed subsets and estimating the variance components for each one. The final variance components are obtained by taking a weighted average of these initial components. The rating method treats the ratings as a random facet, which forces the sparse dataset to be fully crossed. The variance components are then obtained using the same ANOVA procedures that would be used on any other fully crossed dataset. The two methods were first compared in a simulation study to examine the estimation precision under a variety of conditions. There were a total of 27 conditions based on three variables: test taker sample size (*50, 100,* and *200*), number of raters (*4, 8,* and *16*), and rater score variability (*all raters similar score variability, a minority of raters greater score variability than the rest,* and *a majority of raters greater score variability than the rest*). The two methods were then compared in an empirical study to see their effects on the variance component, reliability, and SEM estimates of the MELAB writing section.

The simulation study found that while both methods worked similarly, the rater method outperformed the rating method when the score variability of the raters was more varied. The empirical

3 A version of this report was originally published as: Yu, G. (2007). *Lexical Diversity in MELAB Writing and Speaking Task Performances*, Spaan Working Papers.

study showed that the rating and rater methods were comparable in terms of estimating the variance components, reliability, and SEM for the MELAB writing section. Lin (2014) suggests that this could be a result of the MELAB raters being very well trained in the use of the scoring rubric prior to the rating of operational tests. Additionally, results of the empirical study showed that two raters are sufficient to control measurement error and achieve acceptable score dependability. With two raters we can generally expect to have exact or adjacent agreement since the uncertainty in an awarded score is unlikely to be more than 7.6 scale points (based on a 95% confidence limit). The findings of this study work to provide validity evidence that the number of raters used in the scoring of the MELAB is sufficient to provide test takers with equal opportunities to demonstrate their writing proficiency.

7.5 Performance on the MELAB is related to other indicators of language proficiency in academic contexts

Dobson, Han, and Yamashiro (2001) performed a small-scale concordance study (N = 110) between the MELAB and the TOEFL CBT (CBT). Each participant in the study had taken the CBT within 30 days of their MELAB test date. The average number of days between taking the tests was 16 days. This controlled for any possible confound of language proficiency gains. Approximately 56% of the test takers took the CBT first; the remainder took the MELAB first. This controlled for the possible confound of test order effect. Dobson et al. (2001: 2) noted that the range of scores for the group was "nearly as wide as the range for all who take CBT and MELAB," but that the test takers were, on average, less able than the average test taker for either test:

	μ Study Participants	μ Full Test Population
MELAB	70.00	75.84
TOEFL CBT	181.62	215

The correlation of the participants' scores for each test was 0.89 ($p < 0.01$). Though there was more variability at the lower proficiency levels, a plot of score pairs at each 10th of a percentile (that is, 10th, 20th, 30th, and so on until 90th showed that the relationship is "strongly linear throughout the range of scores (2001: 4). This high correlation suggests that the tests measure English language proficiency in a similar manner. On the basis of this study, Dobson et al. (2001) prepared a concordance table for the total scores on MELAB, TOEFL CBT, and TOEFL PBT. This table has since been augmented with concordances with the TOEFL iBT and is available on the CaMLA website. Nevertheless, Dobson et al. (2001: 1) caution against "translating" the scores from the MELAB into their corresponding TOEFL CBT scores. They point out that the tests are different in format and content and cannot, therefore, be assumed to be measuring the same construct. They suggest that the study provides helpful information for admissions officers who routinely interpret score reports for both tests but that local validation studies should be conducted to confirm the appropriacy of these initial guidelines. Studies of MELAB score use in both academic and professional contexts would provide a useful complement to this work.

7.6 The MELAB has positive consequences for stakeholders

One of the properties of a test that has positive consequences for test takers is that it "measures the same construct in all relevant populations" (Jiao, 2004: 27). There are a number of different subgroups in the MELAB test population: gender; age groups; first language; and purpose for taking the test. To date, only gender has been investigated. Wang (2006) examined the factor structure of the MELAB across gender. Wang's (2006) dataset comprised 216 test takers (68% female). Even though the male test taker group had a lower 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.

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

- The different item types and tasks are representative of the kinds of input and output that students encounter in college and university settings.

- The different item types and tasks are appropriate for measuring intermediate toadvanced levels of language proficiency
- The language elicited by the speaking and writing sections of the MELAB reflects the language expected at the intermediate to advanced levels of language proficiency.
- The language elicited by the speaking and writing sections of the MELAB reflects the language used in college and university settings
- MELAB test results are well understood by test users and are used appropriately.

8. References

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

Aryadoust, V. and Goh, C. C. M. (2013) Exploring the relative merits of cognitive diagnostic models and confirmatory factor analysis for assessing listening comprehension, in Galaczi, E. and Weir, C. J. (Eds.) *Exploring language frameworks: proceedings of the ALTE Kraków Conference, July 2011*, Cambridge: Cambridge University Press, pp. 405–426.

Aryadoust, V., Goh, C. C. M., and Lee, O. K. (2011) An investigation of differential item functioning in the MELAB listening test, *Language Assessment Quarterly*, 8: 361–385.

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

Chapman, M. (2012). *The challenges of ensuring task equivalence in writing tests: A new approach*. Proceedings of the Korea English Language Testing Association: KELTA 2012 Annual International Conference. Seoul, Korea.

Council of Europe (2001) *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.

Dobson, B., Han, I., and Yamashiro, A. D. (2001) *The relationship between test scores on the MELAB and the TOEFL CBT*, UMELIRR2001-01, Ann Arbor, MI: University of Michigan.

Eom, M. (2008) *Underlying factors of MELAB listening constructs*, Spaan Working Papers.

Field, A. (2005) *Discovering Statistics Using SPSS,* London: Sage Publications Inc.

Gao, L. (2006) Toward a cognitive processing model of MELAB reading test item performance, Spaan Working Papers.

Gao, L. and Rogers, W. T. (2011) Use of tree-based regression in the analyses of L2 reading test items, *Language Testing*, 28(1): 77–104.

Goh, C. and Aryadoust, V. S. (2010) *Investigating the construct validity of the MELAB listening test through the Rasch Analysis and Correlated Uniqueness Modeling,* Spaan Working Papers.

Hamp-Lyons, L. and Davies, A. (2006) Bias revisited, Spaan Working Papers.

Hattie, J. (1985) Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9, 139–164.

Jiao, H. (2004) *Evaluating the dimensionality of the Michigan English Language Assessment Battery*, Spaan Working Papers.

Johnson, J. and Lim, G. (2009) The influence of rater language background on writing performance assessment, *Language Testing*, 26(4): 485–505.

Jung, Y. J., Crossley, S. A., and McNamara, D. S. (2014) *Linguistic features in MELAB writing task performances*, Poster presented at the meeting of the American Association for Applied Linguistics (AAAL), Portland, Oregon, March 2014.

Li, H. (2011) *A cognitive diagnostic analysis of the MELAB listening test*, Spaan Working Papers.

Lim, G. S. (2009) *Prompt and rater effects in second language writing performance assessment*. Unpublished PhD dissertation, University of Michigan.

Lim, G. (2010) *Investigating prompt effects in writing performance assessment*, Spaan Working Papers.

Lim, G. (2011) The development and maintenance of rating quality in performance writing assessment: a longitudinal study of new and experience raters, *Language Testing*, 28(4): 543–560.

Lin, C-K. (2014) *Treating either ratings or raters as a random facet in performance-based language assessments: does it matter?* CaMLA Working Papers, 2014-01.

Park, T. (2006) *Detecting DIF across different language and gender groups in the MELAB essay test using the logistic regression method*, Spaan Working Papers.

Purpura, J. M. (1999) *Learner strategy use and performance on language tests: A structural equation modeling approach.* Cambridge, UK: Cambridge University Press.

Purves, A. C., Soter, A., Takala, S., and Vahapassi, A. (1984) Towards a domain-referenced system for classifying composition assignments, *Research in the Teaching of English*, 18: 385–416.

Song, X. (2005) *Language learner strategy use and English proficiency on the Michigan English Language Assessment Battery*, Spaan Working Papers.

Song, X. (2010) Chinese test-takers' performance and characteristics on the Michigan English Language Assessment Battery, in Cheng, L. and Curtis, A. (Eds.) *English language assessment and the Chinese learner*, New York and London: Routledge, ch. 9.

Spaan, M. (1989) Essay tests: what's in a prompt?, paper presented at the 23rd annual convention of Teachers of English to Speakers of Other Languages (TESOL), San Antonio, TX, 7–11, March 1989.

Spaan, M. (1993) The effect of prompt in essay examinations, in D. Douglas and C. Chapelle (Eds.) *A new decade of language testing research: selected papers from the 1990 Language Testing Research Colloquium*, Alexandria, VA: TESOL, Inc., pp. 98–122.

Vahapassi, A. (1982) On the specification of the domain of written composition, in A. Purves and S. Takala (Eds.) *An international perspective on the evaluation of written composition: Evaluation in education: An international review series*, Oxford: Pergamon Press, pp. 265–290.

Wang, S. (2006) *Validation and invariance of factor structure of the ECPE and MELAB across gender*, Spaan Working Papers.

Yu, G. (2009) Lexical Diversity in Writing and Speaking Task Performances, *Applied Linguistics*, 31(2): 236–259