Linking the Common European Framework of Reference and the Michigan English Language Assessment Battery

Technical Report

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

04/2017

TABLE OF CONTENTS

LIST OF TABLES

LIST OF TABLES

1. INTRODUCTION

1.1 OVERVIEW

This report summarizes the results of a multi-panel standard setting study that was conducted with panelists in the United States (US) and the United Kingdom (UK). The purpose of the study was to link scores on each section of the Michigan English Language Assessment Battery (MELAB) to the proficiency levels of the Common European Framework of Reference. This study utilized the Council of Europe's (2009) manual supporting standard setting and Tannenbaum and Cho's (2014) article on critical factors to consider in standard setting as guidelines. This report documents the standard setting study and provides validity evidence to support its quality.

1.2 COMMON EUROPEAN FRAMEWORK OF REFERENCE

The Common European Framework of Reference (CEFR) provides a common basis for evaluating the ability level of language learners. The framework describes "what language learners have to learn to do in order to use a language for communication and what knowledge and skills they have to develop so as to be able to act effectively" (Council of Europe, 2001, p. 1). The CEFR defines six main proficiency levels: A1 and A2 (basic users), B1 and B2 (independent users), and C1 and C2 (proficient users). The CEFR is widely used by test developers and other stakeholders to assist with score interpretation and decision making, so linking the MELAB to the CEFR is beneficial to test users; it will help them to better interpret the test results.

1.3 STANDARD SETTING

Standard setting can be defined as the process of identifying minimum test scores that separate one level of performance from another (Cizek & Bunch, 2007; Tannenbaum, 2011). These minimum test scores, often referred to as cut scores, are defined as the points on a score scale that act as boundaries between adjacent performance levels (Cohen, Kane, & Crooks, 1999). The final product of any standard setting study is the recommended cut scores that link the scores on the test to the target standards or performance descriptors.

The most important component of the standard setting process is the standard setting meeting. During this meeting, facilitators guide a panel of experts through the process of determining cut scores. After a brief introduction to the test and standards in question, the panelists proceed to the first stage of the standard setting meeting, known as familiarization. The purpose of the familiarization stage is to ensure that the panelists understand the standards and performance descriptors to which the test is being linked. The second stage of the standard setting meeting, training, allows the panelists to practice making judgments to ensure that they understand the procedure. During the final stage, judgment, panelists make their individual cut score recommendations. Typically, there are two or more rounds of judgment so that the panelists can discuss their individual decisions, and, if necessary, make adjustments.

Once the standard setting meeting has concluded, the standard setting meeting and the recommended cut scores are examined for procedural, internal, and external validity (Council of Europe, 2009, Ch. 7; Tannenbaum & Cho, 2014). Procedural validity evidence shows that the study plan was implemented as intended, and internal validity evidence shows that the judgments were consistent (Tannenbaum & Cho, 2014). External validity evidence refers to any independent evidence that supports the outcomes of the current study (Council of Europe, 2009, Ch. 7).

1.4 THE MICHIGAN ENGLISH LANGUAGE ASSESSMENT BATTERY

The Michigan English Language Assessment Battery (MELAB) is a standardized English-as-a-foreign-language examination developed and produced by Cambridge Michigan Language Assessments (CaMLA). It is designed to evaluate the English language competence of adult nonnative speakers of English who will need to use English for academic or professional purposes. That being the case, the MELAB is aimed primarily at the B2 (upper intermediate) and C1 (lower advanced) levels, but also measures at the B1 level.

Of the four language skills, the listening, GCVR (grammar, cloze, vocabulary, and reading), and writing sections of the MELAB are required for all test takers, while the speaking section is optional. The listening and GCVR sections consist of several types of multiple choice questions. The listening section has three parts: short recorded questions, short recorded conversations, and recorded interviews. The GCVR section has four parts: grammar questions, cloze passages, vocabulary questions, and reading passages. The writing and speaking sections are constructed response tasks. The writing section asks test takers to write an argumentative essay based on one of two topics, and the speaking section asks test takers to engage in a semi-structured interview with an examiner.

CaMLA is committed to excellence in its tests, which are developed in accordance with the highest standards in educational measurement. All parts of the examination are written following specified guidelines, and items are pretested to ensure that they function properly. CaMLA works closely with test centers to ensure that its tests are

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. METHODOLOGY

2.1 PANEL DESIGN

Standard setting is often described as "fundamentally, a decision-making process" (Skorupski, 2012, p. 135). The "decision-making" aspect is why expert judges are an essential element to successful standard setting, and they become even more important when the performance descriptors in question are from an internationally used framework such as the CEFR. One of the CEFR's biggest strengths (and reason for existence) is its applicability across different contexts. However, some researchers have raised questions about the degree of agreement that there is in the field about what it means for learners across those different contexts to be at a particular level of the CEFR (e.g., de Jong, J. H. A. L., 2013). The question of agreement or lack of agreement seems particularly acute when tests that have similar purposes and are assessing similar constructs do not demonstrate comparable results in terms of CEFR levels when examined through correlations (Lim, Geranpayeh, Khalifa, & Buckendahl, 2013). The contexts of standard setting meetings have been proposed as a possible source of variation (Lim et al., 2013), or in some cases, as an explanation for why cut score decisions were adjusted (Papageorguiou, Tannenbaum, Bridgeman, & Cho, 2015). Therefore, in order to obtain the best possible cut scores, it was decided to hold standard setting meetings in two different contexts, the US and UK, to reflect the US origin of the text and the European origin of the CEFR, and to try to account for this potential variation.

2.2 PANELISTS

As mentioned above, one of the most important features of a standard setting study is the panel of experts that make judgments on the location of the cut scores. It is important that the participants have good knowledge of the examination in question, the test-taking population, and the performance level descriptors (Mills, Melican, & Ahluwalia, 1991; Papageorgiou, 2010). Seven separate panels were conducted for this study, four of which utilized participants from the US, and three smaller ones which utilized participants from the UK. Each of these panels was treated as its own independent linking study. The four US panels each examined one of the four MELAB sections (listening, GCVR, writing, and speaking), and the three UK panels each examined one of the three required[1] MELAB sections (listening, GCVR, and writing).

US Panels

The US-based listening, GCVR, and speaking panels each consisted of thirteen panelists, while the writing panel consisted of fourteen. The majority of the US panelists were recruited from outside CaMLA; however, three panelists on the listening and GCVR panels, four panelists on the writing panel, and one panelist on the speaking panel were selected from CaMLA staff. All of the panelists had experience as ESL/EFL teachers. The speaking panel had an average of more than 9 years of ESL/EFL experience, the GCVR panel had more than 8 years, the writing panel had more than 8 years, and the listening panel had more than 6 years. The listening and GCVR panels also had an average of more than 4 and 5 years of assessment/test development experience, respectively. The writing panel had an average of more than 5 years of writing rater experience, and the speaking panel had an average of more than 4 years of speaking examiner experience. The panelists also had a wide variety of other language testing experience, including experience in test administration, item writing, and scoring. The panelists' experience with standard setting studies and the CEFR prior to the standard setting meeting was varied, so the familiarization activities were particularly important. Overall, the panelists selected for each of the US panels provided a diverse representation of experienced US-based professionals from the field of ESL/EFL.

UK Panels

The UK-based listening panel consisted of five panelists, the UK-based GCVR panel consisted of three panelists, and the UK-based writing panel consisted of four panelists. The UK panelists were all recruited through the Cambridge English's assessment staff and its network of writing examiners and item writers. For the listening and GCVR panels, all of the panelists had experience as ESL/EFL teachers and experience in the field of assessment/test development. The listening panel had an average of more than 11 years of ESL/EFL experience and an average of more than 12 years of assessment/test development experience, while the GCVR panel had an average of more than 17 years of ESL/EFL experience and an average of more than 9 years of assessment/test development experience. While most of the listening and GCVR panelists were quite familiar with the CEFR and standard setting, the familiarization and training activities were still very important. For the

1 Note that a UK panel was not convened for the speaking section due to a number of logistical factors, including the fact that the speaking test is an optional component of the MELAB.

writing panel, all of the panelists were certified writing examiners for the Cambridge English: Advanced (CAE). They all had extensive knowledge of the CAE rating scale, as well as a strong understanding of what features define a C1 level essay. Overall, the panelists selected for each of the UK panels provided a diverse representation of experienced UK-based professionals from the field of ESL/EFL.

2.3 STANDARD SETTING METHOD

There are a variety of standard setting methods in the field of educational measurement. Each method has its own set of advantages and limitations, so the method selected for any study can differ based on many factors, including the type of test involved. This standard setting study primarily utilized two different methods: the Angoff method and the bookmark method.

The Angoff method was first introduced in 1971 and is one of the most widely used procedures for establishing cut scores (Council of Europe, 2009, Ch. 6). This method relies on the concept of a just-qualified or borderline candidate, who can be defined as someone who has only just passed over the threshold between adjacent levels (e.g., a borderline B1/B2 candidate). To make their cut score judgments, panelists must go through the entire test and determine for each item the probability that a just-qualified, borderline candidate would answer it correctly. The test's overall cut score recommendation from each panelist is then calculated by taking the sum of their probability estimates.

The bookmark method is a procedure for establishing cut scores that was developed in 1996 in order to address perceived limitations of other standard setting methods (Cizek, Bunch, & Koons, 2004; Mitzel, Lewis, Patz, & Green, 2001). This procedure is centered on the use of an ordered item booklet, which consists of test items listed in order of increasing difficulty, from the easiest item to the most difficult. The panelists make their cut score judgments by going through the booklet and placing a 'bookmark' at the location where they believe the cut score is located.

US Panels

The US-based standard setting panels applied the Angoff method to the MELAB listening and GCVR sections and the bookmark method to the MELAB writing and speaking sections in order to make three cut score judgments (A2/B1, B1/B2, and B2/C1) for each test section. The Angoff method was selected for the listening and GCVR sections because it allowed us to easily set cut scores on a multiple choice test form, while the bookmark method was selected for the writing and speaking sections because it provided a means of easily setting cut scores on constructed response tasks. Each of the four US panels had two facilitators: one facilitator who served on all four panels and a second facilitator with particular expertise in each of the four MELAB sections who was different for each panel.

For the listening and GCVR sections, the operational items from a previously administered MELAB test form were used for the judgment round test booklets. To make their judgments, the panelists were asked to consider 100 just-qualified candidates at each CEFR level, and state for each item how many of the just-qualified candidates would answer it correctly. This slight modification to the Angoff method is equivalent to asking the panelists to make a probability judgment, but it was done to make it easier for panelists to visualize the task. Due to the time constraints of the standard setting meetings, it was impractical to have the panelists work through the test separately for each target CEFR level. Instead, the panelists were asked to first go through the test section and make their decisions about only the just-qualified B2 level candidates, and then once that was completed, to go through the test section a second time and make their decisions about both the just-qualified B1 and C1 level candidates.

For the writing and speaking sections, the ordered item booklets[2] were created by selecting test taker performances for each possible score point on the rating scales and ordering them from lowest to highest (scores 1–10). Each performance had been scored by at least two certified raters who worked to build a consensus on each performance's score. It should be noted that due to the time constraints of the standard setting meeting, it was impractical to have the panelists listen to the entirety of each speaking performance, so the speaking panel facilitators (who were both certified MELAB speaking test raters) carefully selected audio clips that they determined were most representative of the score awarded for the performance (the clips used were approximately 2- to 3-minute-long excerpts from tests that typically lasted 15 minutes). To make their cut score judgments, the panelists went through the ordered item booklets and placed their bookmarks at the first performance that they felt could have been produced by a just-qualified B1-, B2-, and C1-level candidate.

2 Note that since the speaking performances were audio recordings, the ordered item booklet for the speaking section was actually a digital folder of audio files, not a physical booklet. In practice this digital folder for the speaking section is used in the same way as the physical booklet for the writing section.

UK Panels

For logistical reasons, the UK panels were smaller and somewhat more limited in scope, which in some cases required adjustments to the standard setting approach. For the listening and GCVR sections, the UK-based panels followed the same methodology as the US-based panels. They applied the same standard setting method, the Angoff method, in order to make three cut score judgments (A2/B1, B1/B2, and B2/C1) for each section, and they utilized the same set of materials. The facilitator of the three UK panels was the same facilitator who had helped lead all four US panels.

For the writing section, the UK-based panels utilized a different standard setting method than that of the US panel in order to make a cut score judgment at the level most important to stakeholders and to CAMLA (B2/C1). This panel's participants were asked to do a rating activity where they scored a set of seven MELAB essays (4 essays used in the US-based writing panel [scores 6, 7, 8, & 9] and 3 essays representing midpoint scores not used with the US-based writing panel [scores 6.5, 7.5, & 8.5]) using the CAE writing rating scale, which was already linked to the CEFR. They were also asked to participate in a paired comparison task where they determined whether the seven MELAB essays were better than, similar to, or worse than a CAE essay that had already been rated as a just-qualified C1 performance. The results of these two activities were then used to determine the location of the B2/C1 cut score.

2.4 MEETING PROCEDURES

This section provides an outline of the standard setting meetings for each of the seven panels and summarizes the activities that took place during them. The overall structure of the meetings and the procedures followed during them were generally the same across meetings, though the CEFR scales selected for the familiarization activities (see Appendix A for a list of the scales selected for each test section) and the standard setting method selected for the judgment activity differed slightly. The procedures and results of each standard setting meeting were documented throughout each meeting using Google spreadsheets, and they were analyzed after each meeting to help provide evidence of procedural, internal, and external validity to support the recommended cut scores.

US Panels

Prior to the standard setting meetings, the panelists were required to complete several pre-study activities to begin familiarizing (or, as was the case for many panelists, re-familiarizing) themselves with the MELAB and the CEFR. After completing a brief background questionnaire, the panelists were also asked to complete a pre-study CEFR quiz to assess their understanding of the CEFR prior to the standard setting meetings. This quiz required panelists to assign CEFR levels to 18 descriptors selected from several scales related to the test section being linked. Once the quiz was completed, the panelists were asked to familiarize themselves with the MELAB by reading information on the CaMLA website. They were also asked to familiarize themselves with the CEFR by reading Morrow (2004). Members of all four panels reviewed the CEFR global scale (Council of Europe, 2001, p. 24) and members of the US-based listening, GCVR, and writing panels also reviewed the self-assessment grid (Council of Europe, 2001, p. 26–27); the members of the US-based speaking panel reviewed the table describing qualitative aspects of spoken language use (Council of Europe, 2001, p. 28–29). After reviewing the two CEFR scales assigned for their panel, the panelists were then asked to describe their initial impressions of the characteristics of an average and a just-qualified B1-, B2-, and C1-level candidate. See Appendix B for an example of the pre-study activity questions, which were taken (with some modification), from the Tannenbaum and Wylie (2008) standard setting report.

Each standard setting meeting began with a brief introduction to the standard setting procedure and the goals of the study. The pre-study materials were then reviewed and discussed to address any of the panelists' questions. The discussion primarily focused on the panelists' descriptions of the just-qualified candidates. This helped each panel to understand the characteristics of just-qualified candidates and highlighted their importance.

To familiarize the panelists with the CEFR levels and descriptors, each panel[3] participated in two activities that utilized descriptors from CEFR scales related to the panel's test section. For the first familiarization activity, the panelists began by reviewing and discussing two CEFR scales. The discussion focused on understanding how the descriptors defined each CEFR level, as well as what features a just-qualified B1-, B2-, and C1-level learner would exhibit. After the discussion, the panelists were given a set of descriptors from these scales and were asked to individually assign CEFR levels to each of them. The results were then discussed as a group to help clarify any misclassified descriptors and to ensure that the panelists understood the CEFR

3 A minor scheduling conflict during the US-based speaking panel's meeting resulted in the order of some tasks in the familiarization activities being rearranged. However, this only resulted in a reordering of the tasks; the panelists still completed both familiarization activities, and the results were discussed just as thoroughly as they were for the other panels.

levels. The second familiarization activity was similar to the first; however, it did not include an initial review or discussion of the scales. The panelists began the activity by individually assigning CEFR levels to a set of descriptors from several different scales related to the panel's test section. Because these scales were not discussed prior to the activity, panelists needed to use their knowledge and understanding of the CEFR to help them complete the activity. As before, the results of this activity were then discussed as a group to ensure that the panelists understood the descriptors for each CEFR level. Overall, while the sorting activities utilized by these familiarization activities can be rather challenging due to the decontextualization of the descriptors, they helped to encourage panelist familiarization with the CEFR by forcing them to fully read and deeply consider the language of each descriptor.

The training activity provided the panelists the opportunity to practice making cut score judgments using the Angoff (listening and GCVR panels) or bookmark (writing and speaking panels) method prior to the actual judgment activity. Each panel was provided with the appropriate training materials for the test section: a test booklet with a subset of a MELAB test form's listening items for the listening panel, a test booklet with a subset of a MELAB test form's GCVR items for the GCVR panel, an ordered item booklet of five writing performances for the writing panel, and an ordered item booklet of four speaking performances for the speaking panel. Fewer items were selected for the listening and GCVR training booklets, and a narrower range of performances were selected for the writing and speaking ordered item booklets in order to help reduce the panelists' workload for the training activity, the primary goal of which was to allow panelists to focus on understanding the judgment process. Towards this end, the panelists practiced making their cut score judgments at the B1/B2 boundary for the listening, GCVR, and writing sections, and at the B2/C1 boundary for the speaking section. Once the panelists finished making their practice judgments, each panel discussed the procedures to address any questions or concerns. Once these discussions concluded, the panelists were given a pre-judgment survey to assess their understanding of the procedures and their willingness to proceed with the judgment activity.

For the judgment activity, each panel followed the same procedures that they practiced during the training activity to make their cut score judgments at the A2/B1, B1/B2, and B2/C1 boundaries. The meeting facilitators emphasized the importance of thinking about the just-qualified candidate at each level when making their decisions. Each panel was provided with the appropriate judgment materials for the test section: a test booklet with a MELAB test form's operational listening items for the listening panel, a test booklet with a MELAB test form's operational GCVR items for the GCVR panel, an ordered item booklet of ten writing performances representative of the ten score points on the MELAB writing rating scale for the writing panel, and an ordered item booklet of ten speaking performances representative of the ten score points on the MELAB speaking scale for the speaking panel. The panelists also had access to their notes and the CEFR scales that had been discussed during the familiarization activities.

The judgment activity consisted of two judgment rounds where panelists marked their decisions on spreadsheets. Both judgment rounds were followed by a group discussion of the results. The discussion of the first judgment round allowed panelists to review the items and materials and discuss the reasoning behind their cut score decisions. The panelists reviewed several test items (listening and GCVR panels) and test taker performances (writing and speaking panels) as a group so that they could discuss the factors that influenced their decisions. The listening and GCVR panels were also provided with IRT difficulty statistics for each item to consider during the discussions.

The second judgment round utilized the same materials as the first. The panelists were instructed to perform the judgment activity again, taking into account the discussions of the first judgment round, and, if they felt it was necessary, make adjustments to their cut score decisions. The discussion of the second judgment round focused on finalizing the panel's cut score recommendations. Once the cut score recommendations were finalized, the panelists were given a post-judgment survey to collect their opinions on the quality of the meeting and their confidence in the recommended cut scores, as well as a post-study CEFR quiz to assess how much their knowledge of the CEFR descriptors had improved.

Overall, the procedures and results of the four standard setting meetings were documented throughout each meeting using Google spreadsheets, and they were analyzed after each meting to help provide evidence of procedural, internal, and external validity to support the recommended cut scores.

UK Panels

For the most part, the UK-based listening and GCVR panel meetings followed the same procedures as the US-based panel meetings. The panelists were asked to complete a background questionnaire and review the CEFR global scale and self-assessment grid prior to the meeting, and the meeting itself consisted of a

brief introduction to the MELAB and standard setting, several familiarization activities (two for listening, three for GCVR), a training activity, and two judgment rounds. As in the US panels, each of these activities was followed by an in-depth discussion of the results. The only major difference between the US and UK panels were the familiarization activities. All of the UK panel familiarization activities followed the same format as the first familiarization activities from the US panels. That is, each familiarization activity began with a review and discussion of several CEFR scales, after which, the panelists were given a set of descriptors from these scales and were asked to individually assign CEFR levels to each of them. The results were then discussed as a group to help clarify any misclassified descriptors and to ensure that the panelists understood the CEFR levels. This change was made to the familiarization activities in order to ensure that the panelists had the best possible understanding of the CEFR before the judgment task.

The UK-based writing panel meeting differed from the other meetings. It did not require any familiarization or training activities since the participants were already certified CAE examiners who were simply being asked to use their expertise to rate and compare several essays. Because of this, the meeting was able to be conducted remotely via videoconference. Prior to the meeting, the panelists were asked to complete the rating and paired comparison actives for the MELAB essays. During the meeting the raters discussed their ratings for each essay and explained their reasoning behind their scores. Once the meeting concluded, the raters were asked to do the rating and paired comparison activities again, taking into account the discussions of the essays.

3. RESULTS

3.1 SPECIFICATION

The first stage of a standard setting study, known as specification (Council of Europe, 2009) or construct congruence (Tannenbaum & Cho, 2014), provides evidence that the skills and abilities measured by the test are "consistent with those described by the framework" (Tannenbaum & Cho, 2014, p. 237). This step is often done prior to the standard setting meeting. It requires that the test developers justify the appropriateness of the linking study by showing that the test content is aligned with the target framework. This justification is necessary because, as Tannenbaum and Cho note, "If the test content does not reasonably overlap with the framework of interest, then there is little justification for conducting a standard setting study, as the test would lack content-based validity" (2014, p. 237).

While the MELAB was introduced prior to the development of the CEFR, linking MELAB test scores to the CEFR is justifiable. This justification rests on the understanding that the CEFR was developed as a tool that can describe a broad range of activities, competences, and proficiencies and which can be used with some flexibility (North, 2014). Across the four skill sections of the MELAB, the overlap between the skills and proficiency levels it tests and the activities and proficiencies described in the CEFR scales was deemed sufficient for linking to the CEFR. In terms of the range of language activities specified in the CEFR's illustrative scales, for each MELAB section there were multiple relevant scales (e.g., overall oral production for the speaking section, writing reports and essays for the writing section, understanding conversation between native speakers for the listening section, and overall reading comprehension for the GCVR section; see Appendix A for a full list of the CEFR illustrative scales deemed relevant to the MELAB and used by each panel). It was also sufficient in terms of proficiency levels: the MELAB was specifically designed to assess the English language ability of test takers at lower intermediate to lower advanced levels equivalent to those described by the B1–C1 levels of the CEFR.

3.2 FAMILIARIZATION

This section summarizes the results of the familiarization activities performed during the standard setting meetings for each panel. These activities are important because they help to establish the panelists' familiarity with the CEFR. If panelists did not understand the CEFR levels and their descriptors, then the validity of the recommended cut scores would be jeopardized, since the panelists' judgments may then reflect this lack of understanding.

The results of the familiarization activities for each panel are summarized in the tables in Appendix C. These tables show the number and percentage of descriptors correct, the Spearman correlation (ρ) between the panelists' assigned CEFR levels and the correct descriptor levels, and the average assigned CEFR level for each panelist. The correlation coefficient shows the degree to which the panelists understand the progression of the CEFR levels and should be interpreted in conjunction with the number and percentage of descriptors correct to understand the panelists' performance on the familiarization tasks. The average assigned CEFR level for each panelist was calculated by transforming their assigned CEFR levels to numbers (A1 = 1, A2 = 2, B1 = 3, B2 = 4, C1 = 5, C2 = 6) and taking the average. The panelists' averages can be compared with the average level of the descriptors to assess the overall severity or leniency of the panelists. Panelists with average assigned CEFR

levels higher than the actual average were generally more lenient, while panelists with average assigned CEFR levels lower than the actual averages were generally more severe.

Assigning exact CEFR levels to individual descriptors is a challenging task, but the data presented in Appendix C show that all of the panels performed reasonably well on the familiarization activities. On average, each panel assigned the correct CEFR level to a large percentage of the descriptors (52.7% – 86.3%). Furthermore, analysis of the panelists' individual responses revealed that the vast majority of incorrectly assigned descriptors were placed at adjacent CEFR levels. In addition to the number of correctly assigned descriptors, the relatively high average correlation coefficients for each panel (0.80 – 0.98) also provide evidence that the panelists understood the progression of language proficiency across the different CEFR levels. Finally, the tables show that while the averages of the assigned CEFR levels indicate that the panelists' leniency and severity are varied, as a group they tended to be somewhat lenient. Overall, the results summarized in these tables suggest that the panelists had a very good understanding of the CEFR descriptors. This understanding was strengthened through group discussion of the descriptor statements following each familiarization activity. These discussions were held to correct any misunderstandings and to ensure that the panelists understood the correct CEFR level for each descriptor.

In addition to analyzing the panelists' individual understandings of the descriptors, it is also important when examining panelist familiarity with the CEFR to assess the consistency of each panel as a whole since the cut scores will be based on each panel's decisions. Table 3.1 presents three measures of internal consistency for each panel's familiarization activities: Cronbach's alpha (α), the intraclass correlation coefficient (ICC), and Kendal's coefficient of concordance (W). These indices are three of the most frequently used measures of internal consistency (Kaftandjieva, 2010, p. 96). Cronbach's alpha (α) measures internal consistency by estimating the proportion of variance due to common factors in the items (Davies et al., 1999, p. 39), the ICC measures internal consistency by taking into account both between- and within-rater variance (Davies et al., 1999, p. 89), and Kendall's W is a nonparametric measure of internal consistency that measures the level of agreement between three or more raters that rank the same group of items (Davies et al., 1999, p. 100). These three indices range from 0 to 1, with a value of 1 indicating complete agreement among panelists. Table 3.1 shows that all three indices were very high, with Cronbach's alpha (α) and ICC values very close to 1 for all panels. This suggests that there was a very high level of agreement and consistency between the panelists for each of the four panels.

Table 3.1: Panel Agreement and Consistency for Familiarization Activities

Panel	Activity	α	ICC*	W
Listening (US)	1	0.990	0.990	0.857
	2	0.984	0.983	0.806
Listening (UK)	1	0.986	0.986	0.914
	2	0.988	0.988	0.903
GCVR (US)	1	0.991	0.991	0.878
	2	0.989	0.989	0.861
GCVR (UK)	1	0.981	0.981	0.936
	2	0.970	0.970	0.898
	3	0.982	0.979	0.932
Writing (US)	1	0.981	0.982	0.718
	2	0.987	0.986	0.833
Speaking (US)	1	0.990	0.990	0.864
	2	0.987	0.985	0.832

* ICC values obtained using a two-way mixed model and average measures for exact agreement.

The familiarization activities are meant to expose panelists to the CEFR descriptors relevant to the study and ensure that they all had an accurate understanding of each CEFR level. While the above analysis demonstrates that the panelists had a good understanding of the CEFR descriptors, it is important to note that these were learning activities, so some inaccuracies and inconsistencies from the panelists were expected at this stage. The descriptor statements were thoroughly discussed after each familiarization task, and any questions on the levels of the descriptor statements were addressed to ensure that the panelists understood the correct level of each descriptor.

One measure of the effectiveness of the familiarization tasks can be obtained through analysis of the pre- and post-study CEFR quizzes. Per Section 2.3, the US and UK panelists were all given a short CEFR quiz with their pre-study materials to assess their initial understanding of the CEFR and another version of this quiz at the conclusion of the study to assess whether their understanding of the CEFR had improved. Tables 3.2–3.7 summarize the results of both quizzes for each panel (reported as raw number correct from a total of 18 descriptors). They reveal that, on average, the panelists' scores improved for each panel after the standard setting meeting. Analysis of each panel's data with a paired t-test confirmed that this positive difference in scores was statistically significant for the US listening ($t=2.19$, $df=12$, $p=0.049$), US GCVR ($t=3.33$, $df=12$, $p=0.006$), and US speaking ($t=2.56$, $df=12$, $p=0.025$) panels, but not for the amount of improvement

Table 3.2: US Listening Panel Pre- and Post-Study CEFR Quiz Results (number correct from 18 total)

Panelist ID	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13	Average	SD
Pre-Study	9	11	11	9	9	13	10	3	7	11	12	13	11	9.92	2.69
Post-Study	7	13	13	9	13	14	14	12	13	10	13	10	14	11.92	2.22
Difference	-2	2	2	0	4	1	4	9	6	-1	1	-3	3	2.00	-0.47

Table 3.3: US GCVR Panel Pre- and Post-Study CEFR Quiz Results (number correct from 18 total)

Panelist ID	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	Average	SD
Pre-Study	9	13	10	15	11	1	9	14	9	11	13	12	10	10.54	3.48
Post-Study	13	15	14	10	14	7	13	16	13	12	14	14	14	13.00	2.31
Difference	4	2	4	-5	3	6	4	2	4	1	1	2	4	2.46	-1.17

Table 3.4: US Writing Panel Pre- and Post-Study CEFR Quiz Results (number correct from 18 total)

Panelist ID	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14	Average	SD
Pre-Study	10	2	12	10	4	15	11	11	11	10	15	12	13	8	10.29	3.65
Post-Study	14	7	10	9	7	14	12	10	11	11	16	13	13	9	11.14	2.68
Difference	4	5	-2	-1	3	-1	1	-1	0	1	1	1	0	1	0.85	-0.97

Table 3.5: US Speaking Panel Pre- and Post-Study CEFR Quiz Results (number correct from 18 total)

Panelist ID	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	Average	SD
Pre-Study	15	6	15	12	6	11	12	14	7	6	13	11	11	10.69	3.38
Post-Study	15	10	13	16	14	13	12	12	13	11	12	13	14	12.92	1.61
Difference	0	4	-2	4	8	2	0	-2	6	5	-1	2	3	2.23	-1.77

Table 3.6: UK Listening Panel Pre- and Post-Study CEFR Quiz Results (number correct from 18 total)

Panelist ID	L1	L2	L3	L4	L5	Average	SD
Pre-Study	13	11	13	12	13	12.40	0.89
Post-Study*	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Difference	N/A	N/A	N/A	N/A	N/A	N/A	N/A

*Due to time limitations, the post-study quiz was not able to be administered for this panel.

Table 3.7: UK GCVR Panel Pre- and Post-Study CEFR Quiz Results (number correct from 18 total)

Panelist ID	S1	S2	S3	Average	SD
Pre-Study	12	13	10	11.67	1.53
Post-Study	16	14	13	14.33	1.53
Difference	4	1	3	2.67	0.00

demonstrated by the UK GCVR (t=3.02, df=2, p=0.094) and US writing (t=1.61, df=13, p=0.132) panels. These results provide evidence that the familiarization activities and their discussions helped to improve the panelists' understanding of the CEFR descriptors.

Overall, the analysis of the familiarization activities reveals that the panelists had a good understanding of the CEFR levels and that the activities and discussions were successful in helping them understand the CEFR descriptors. The comments made throughout the discussion of the familiarization activities, the responses to the pre- and post-judgment surveys (see Section 4.1), and the low variability of the judgment task (see Section 3.3) also suggest that the panelists understood the CEFR levels and the differences between adjacent levels.

3.3 JUDGMENT

This section summarizes the results of the judgment activities. Tables 3.8–3.15, below, present the results of these activities for each panel. The tables provide each panelist's individual cut score recommendations as well as summary statistics for the panel as a whole for both judgment rounds. Of particular interest are the average cut scores, which represent the panels' initial cut score recommendations for each section of the MELAB.

Listening

Tables 3.8 and 3.9 summarize the results of the judgment activities for the US and UK listening panels (37 total items were judged). They show that the panelists' cut score recommendations were all quite similar within each panel, and that there was little variation in the panelists' individual cut score recommendations for each level. After discussing the results of the second judgment round, the US panel decided that an A2/B1 cut score of 12, a B1/B2 cut score of 24, and a B2/C1 cut score of 33 were most representative of their cut score recommendations, and the UK panel decided that an A2/B1 cut score of 14, a B1/B2 cut score of 23, and a B2/C1 cut score of 31 were most representative of their cut score recommendations. These initial cut score recommendations were then averaged together to determine the final raw cut scores for the MELAB listening section. This resulted in an A2/B1 cut score of 13, a B1/B2 cut score of 24, and a B2/C1 cut score of 32.

Table 3.8: US Listening Panel Cut Score Judgments

Panelist ID	Judgment Round 1			Judgment Round 2		
	A2/B1	B1/B2	B2/C1	A2/B1	B1/B2	B2/C1
L1	3.90	21.55	35.63	7.50	24.35	35.92
L2	9.00	26.65	34.48	9.25	26.95	34.51
L3	13.10	26.80	33.20	11.70	24.85	33.25
L4	22.10	26.65	36.03	19.50	26.83	36.02
L5	5.25	17.30	31.80	11.15	23.55	33.35
L6	11.68	24.50	33.27	11.59	24.80	33.33
L7	17.60	24.90	33.95	16.30	24.90	33.85
L8	24.00	28.40	33.50	19.70	26.50	32.35
L9	8.41	22.70	33.54	10.24	23.95	33.38
L10	8.35	18.00	32.00	8.50	19.30	31.80
L11	10.95	25.70	31.75	11.15	25.45	31.20
L12	5.45	24.70	33.05	7.60	24.90	33.15
L13	6.50	20.10	30.00	7.65	20.80	30.45
Average	11.25	23.69	33.25	11.68	24.39	33.27
Median	9.00	24.70	33.27	11.15	24.85	33.33
SD	6.40	3.51	1.63	4.24	2.21	1.63
Min	3.90	17.30	30.00	7.50	19.30	30.45
Max	24.00	28.40	36.03	19.70	26.95	36.02

Table 3.9: UK Listening Panel Cut Score Judgments

Panelist ID	Judgment Round 1			Judgment Round 2		
	A2/B1	B1/B2	B2/C1	A2/B1	B1/B2	B2/C1
L1	10.60	20.00	32.85	11.45	20.20	31.95
L2	9.62	21.98	29.89	10.23	22.43	30.28
L3	14.45	23.38	32.01	14.47	23.44	31.92
L4	17.52	23.70	30.32	18.10	23.90	30.67
L5	17.60	25.90	27.90	17.65	25.40	28.00
Average	13.96	22.99	30.59	14.38	23.07	30.56
Median	14.45	23.38	30.32	14.47	23.44	30.67
SD	3.75	2.18	1.93	3.55	1.93	1.61
Min	9.62	20.00	27.90	10.23	20.20	28.00
Max	17.60	25.90	32.85	18.10	25.40	31.95

GCVR

Table 3.10 and 3.11 summarize the results of the judgment activities for the US and UK GCVR panels (65 total items were judged). They show that the panelists' cut score recommendations were all quite similar within each panel, and that there was little variation in the panelists' individual cut score recommendations for each level. After discussing the results of the second judgment round, the US panel decided that an A2/B1 cut score of 23, a B1/B2 cut score of 42, and a B2/C1 cut score of 59 were most representative of their cut score recommendations, and the UK panel decided that an A2/B1 cut score of 16, a B1/B2 cut score of 32, and a B2/C1 cut score of 46 were most representative of their cut score recommendations. These initial cut score recommendations were then averaged together to determine the final raw cut scores for the MELAB GCVR section. This resulted in an A2/B1 cut score of 20, a B1/B2 cut score of 37, and a B2/C1 cut score of 52.

Table 3.11: UK GCVR Panel Cut Score Judgments

Panelist ID	Judgment Round 1			Judgment Round 2		
	A2/B1	B1/B2	B2/C1	A2/B1	B1/B2	B2/C1
R1	15.70	28.60	35.95	16.15	29.40	39.40
R2	10.80	20.75	45.95	14.05	31.05	50.20
R3	18.20	35.75	50.95	17.80	34.40	49.50
Average	14.90	28.37	44.28	16.00	31.62	46.37
Median	15.70	28.60	45.95	16.15	31.05	49.50
SD	3.76	7.50	7.64	1.88	2.55	6.04
Min	10.80	20.75	35.95	14.05	29.40	39.40
Max	18.20	35.75	50.95	17.80	34.40	50.20

Table 3.10: US GCVR Panel Cut Score Judgments

Panelist ID	Judgment Round 1			Judgment Round 2		
	A2/B1	B1/B2	B2/C1	A2/B1	B1/B2	B2/C1
R1	26.55	42.56	59.01	25.17	41.64	58.95
R2	15.35	45.30	58.45	15.75	44.70	58.40
R3	25.75	47.44	62.86	25.55	47.30	62.78
R4	11.65	33.90	57.77	18.09	40.28	57.73
R5	21.82	42.04	59.22	22.37	42.24	59.39
R6	34.59	44.83	55.69	28.28	43.23	57.11
R7	33.95	46.85	58.35	30.50	45.65	57.90
R8	11.60	33.10	58.36	14.20	35.65	59.01
R9	24.70	38.55	60.43	23.95	41.55	60.79
R10	18.00	39.50	56.19	17.70	39.45	54.80
R11	30.15	44.05	56.10	25.70	40.75	56.55
R12	30.25	45.55	61.98	30.80	46.60	62.22
R13	22.35	40.00	55.85	22.45	39.85	55.85
Average	23.59	41.82	58.48	23.12	42.22	58.58
Median	24.70	42.56	58.36	23.95	41.64	58.40
SD	7.75	4.62	2.28	5.38	3.25	2.35
Min	11.60	33.10	55.69	14.20	35.65	54.80
Max	34.59	47.44	62.86	30.80	47.30	62.78

Table 3.12: US Writing Panel Cut Score Judgments

Panelist ID	Judgment Round 1			Judgment Round 2		
	A2/B1	B1/B2	B2/C1	A2/B1	B1/B2	B2/C1
W1	3	6	9	3	6	9
W2	3	6	9	4	6	9
W3	3	6	9	3	6	9
W4	3	6	9	3	6	9
W5	4	6	9	4	6	9
W6	5	7	9	4	7	9
W7	4	6	8	4	6	9
W8	3	6	9	3	6	9
W9	3	6	9	3	6	9
W10	4	5	7	4	6	9
W11	3	6	9	3	6	9
W12	4	6	8	4	7	9
W13	4	7	9	4	7	9
W14	3	7	9	3	6	9
Average	3.50	6.14	8.71	3.50	6.21	9.00
Median	3	6	9	3.5	6	9
SD	0.65	0.53	0.61	0.52	0.43	0.00
Min	3	5	7	3	6	9
Max	5	7	9	4	7	9

Writing

Table 3.12 summarizes the results of the judgment activities for the US writing panel (ten total writing performances were judged). It shows that the US panelists' cut score recommendations were all very similar, and that the panelists even had unanimous agreement on the B2/C1 cut score. After discussing the results of the second judgment round, the US panel decided that an A2/B1 cut score of 4, a B1/B2 cut score of 6, and a B2/C1 cut score of 9 were most representative of their cut score recommendations. However, during this discussion several panelists voiced concerns that there was such a large jump in test taker ability level between the essay with an 8 and the essay with a 9 that the performance of a just-qualified C1 test taker might fall somewhere between the two scores.

Tables 3.13 and 3.14 summarize the results of the judgment activities for the UK writing panel (seven total writing performances were judged). This panel made judgments, using the CAE rating scale, on writing performances that included several midpoint scores not used with the US panel to determine if they provided a more appropriate location for the B2/C1 cut score. Table 3.13 shows that the raters only awarded C1 level scores (CAE scores of 9 or higher) to the MELAB essays with a score of 8.5 and 9. None of the other essays were scored at the C1 level by any of the four raters. Table 3.14 shows that the paired comparison activity confirms these results since the raters ranked the essays scored 6 through 8 as worse than a just-qualified C1 essay from the CAE, and essays scored as 8.5 and 9 were ranked as similar to or better than the just-qualified C1 essay. Overall, the results of the UK panel suggest that a score of 8.5 corresponds to the B2/C1 cut score. The results of the US and UK panels were then combined, which resulted in an A2/B1 cut score of 4, a B1/B2 cut score of 6, and a B2/C1 cut score of 8.5.

Table 3.13: UK Writing Panel Rating Activity

Rater ID	Round 1							Round 2						
	6*	6.5*	7*	7.5*	8*	8.5*	9*	6*	6.5*	7*	7.5*	8*	8.5*	9*
W1	8	1	5	5	5	14	9	5	3	5	6	6	12	10
W2	4	5	10	6	8	11	10	4	5	6	6	7	11	10
W3	7	7	8	8	8	8	9	6	8	7	8	7	8	10
W4	3	2	7	6	5	9	12	3	3	8	6	5	9	11
Average	5.50	3.75	7.50	6.25	6.50	10.50	10.00	4.50	4.75	6.50	6.50	6.25	10.00	10.25
Median	5.50	3.50	7.50	6.00	6.50	10.00	9.50	4.50	4.00	6.50	6.00	6.50	10.00	10.00
SD	2.38	2.75	2.08	1.26	1.73	2.65	1.41	1.29	2.36	1.29	1.00	0.96	1.83	0.50
Min	3	1	5	5	5	8	9	3	3	5	6	5	8	10
Max	8	7	10	8	8	14	12	6	8	8	8	7	12	11

*Note that the essays marked 6, 7, 8, & 9 represent MELAB scores used with the US panel, and essays marked 6.5, 7.5, & 8.5 represent midpoint MELAB scores not used with the US panel.

Table 3.14: UK Writing Panel Paired Comparison Activity

Rater ID	Better than the Just-Qualified C1 Essay	Similar to the Just-Qualified C1 Essay	Worse than t he Just-Qualified C1 Essay
W1	8.5	9	6, 6.5, 7, 7.5, 8
W2	-	8.5	6, 6.5, 7, 7.5, 8
W3	-	9	6, 6.5, 7, 7.5, 8, 8.5
W3	9	8.5	6, 6.5, 7, 7.5, 8

Speaking

Table 3.15 summarizes the results of the judgment activities for the speaking panel (ten total speaking performances were judged). It shows that the panelists' cut score recommendations were all very similar, and that the panelists even had unanimous agreement in judgment round two on the A2/B1 and B1/B2 cut scores. Due to the high amount of agreement in this panel (only one panelist recommended a different cut score), there was very little to discuss from the second judgment round. The panel decided that an A2/B1 cut score of 3, a B1/B2 cut score of 6, and a B2/C1 cut score of 9 were most representative of the panel's cut score recommendations.

Table 3.15: US Speaking Panel Cut Score Judgments

Panelist ID	Judgment Round 1			Judgment Round 2		
	A2/B1	B1/B2	B2/C1	A2/B1	B1/B2	B2/C1
S1	3	6	9	3	6	9
S2	3	6	8	3	6	9
S3	4	7	9	3	6	9
S4	3	7	8	3	6	9
S5	4	6	9	3	6	9
S6	3	6	9	3	6	9
S7	3	6	9	3	6	9
S8	3	6	8	3	6	9
S9	3	6	9	3	6	9
S10	3	8	10	3	6	9
S11	3	6	8	3	6	8
S12	3	6	8	3	6	9
S13	3	6	8	3	6	9
Average	3.15	6.31	8.62	3.00	6.00	8.92
Median	3	6	9	3	6	9
SD	0.38	0.63	0.65	0.00	0.00	0.28
Min	3	6	8	3	6	8
Max	4	8	10	3	6	9

4. VALIDITY EVIDENCE

4.1 PROCEDURAL VALIDITY

The documentation of the standard setting study throughout this report provides procedural validity evidence to support the quality of the standard setting meetings and the cut score recommendations. This section provides additional procedural validity evidence by summarizing the panelists' responses to pre- and post-judgment surveys that were given during the standard setting meetings. The pre-judgment survey focused on the panelists' understanding of the familiarization and training activities, while the post-judgment survey focused on the panelists' understanding of the judgment rounds and their confidence in the recommended cut scores. Both surveys used a four-point Likert scale (1 – strongly disagree to 4 – strongly agree) to collect most of this information. Tables 4.1 and 4.2 present the statements and summarize the results for the pre- and post-judgment surveys, respectively. In addition to the statements listed in the tables, the pre-judgment survey asked the panelists to indicate if they were ready to proceed to the judgment task (yes or no), and the post-judgment survey asked the panelists to indicate their opinion of the recommended cut scores (too low, about right, or too high).

These tables show that the panelists generally responded favorably to the survey statements across all panels. The majority of panelists indicated that they understood the familiarization, training, and judgment activities and expressed confidence in their decisions and indicated that they had enough time to complete their tasks and participate in group discussions. On the pre-judgment survey, all of the panelists indicated that they

Table 4.1: Summary of Pre-Judgment Survey Results

No.	Statement	Listening*				GCVR*				Writing				Speaking			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	The familiarization activities helped me to understand the CEFR levels	-	-	6	12	-	-	2	14	-	-	4	10	-	-	2	11
2	The training activity helped me to understand the judgment process	1	-	5	12	-	-	1	15	-	-	3	11	-	-	2	11
3	I had enough time to complete my individual tasks	-	1	4	13	-	-	2	14	-	-	-	14	-	-	1	12
4	I had enough time to participate in the discussions	-	-	1	17	-	-	1	15	-	-	2	12	-	-	1	12

*Note that the summary for the Listening and GCVR panels include the responses from both the US and UK panels.

Table 4.2: Summary of Post-Judgment Survey Results

No.	Statement	Listening*				GCVR*				Writing				Speaking			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	The familiarization activities helped me to understand the CEFR levels	-	-	4	14	-	-	1	15	-	-	2	12	-	-	2	11
2	The training activity helped me to understand the judgment process	-	-	3	15	-	-	-	16	-	-	2	12	-	-	1	12
3	I understood the instructions for each judgment round.	-	-	3	15	-	-	-	15**	-	-	1	13	-	-	-	13
4	I understood the group discussion of our judgments	-	-	-	18	-	-	-	16	-	-	-	14	-	-	1	12
5	I had enough time to complete my individual tasks	-	1	2	15	-	1	2	13	-	-	-	14	-	-	1	12
6	I had enough time to participate in the discussions	-	-	1	17	-	-	1	15	-	-	1	13	-	-	-	13
7	I am confident in the decisions I have made	-	1	7	10	-	-	6	10	-	-	1	13	-	-	3	10

*Note that the summary for the Listening and GCVR panels include the responses from both the US and UK panels.

**Note that one panelist elected to write in his/her own score (2.5) for GCVR statement 3, rather than select one of the possible score points.

felt ready to continue to the judgment activity, and on the post-judgment survey, all of the panelists indicated that the cut score recommendations were about right. Out of the 61 panelists that participated in these linking activities, only four disagreed with any of the survey statements.

One panelist, from the UK listening panel, disagreed with pre-judgment statement 3, indicating that she did not feel as though she had enough time to complete her individual tasks. However, she noted that this was only due to the internet connectivity issues that she had experienced. Another panelist, also from the UK listening panel, disagreed with pre-judgment statement 2, indicating that she did not feel the training activity helped her to understand the judgment process. However, this panelist also indicated agreement with this same statement on the post-judgment survey, which suggests that the panelist may have just had some initial discomfort with the Angoff method that was resolved during the judgment rounds, rather than a lack of understanding. Two panelists, one from the US GCVR panel and one from the US listening panel, disagreed with post-judgment statement 5, indicating that they felt that they did not have enough time to complete their individual tasks. The same panelist from the US listening panel also disagreed with post-judgment statement 7, indicating that she did not feel confident in the decisions she made. It is possible that the difficulty of the Angoff method may have contributed to the panelists' concerns. However, despite these responses, the panelists' responses to the other survey statements indicate that they understood the standard setting procedure and that they thought the recommended cut scores were appropriate. This suggests that while these panelists may have lacked confidence in their decisions or felt that they needed more time to complete their tasks, they still understood the procedures and felt that the panel arrived at appropriate cut scores.

Overall, the generally positive responses to the pre- and post-judgment surveys indicate that, as a whole, the panelists understood the standard setting procedure and were satisfied with the cut score recommendations. This provides procedural validity evidence that supports the quality of the cut score recommendations.

4.2 INTERNAL VALIDITY

This section provides internal validity evidence to support the recommended cut scores for each section of the MELAB. One piece of internal validity evidence can be obtained by examining the likelihood that the recommended cut scores from each panel can be replicated. This can be estimated using the standard error of judgment (SEj) of each panel's cut score recommendations (Tannenbaum & Cho, 2014). Cohen, Kane, and Crooks (1999) suggest that SEj values that are less than half the test's standard error of measurement (SEM) can be considered reasonable. That is, if the SEj values are less than half the test's SEM, then the recommended cut scores would likely be replicated in another standard setting study.

SEM estimates for the listening, GCVR, and writing sections were obtained using the 2015 MELAB test data, and SEM estimates for the speaking section were

obtained using examiner monitoring data[4]. The panelists' judgments for the listening, GCVR, and writing panels were not originally made using MELAB scaled scores, so their raw cut score recommendations needed to be transformed onto the appropriate scale before we could calculate the SEj values for comparison. This was done by rounding the panelists' cut score recommendations to the nearest whole number and applying the appropriate raw-to-scale conversion table. Table 4.3 presents the SEj values for each panel's cut scores[5], as well as the SEM estimates for each test section.

Table 4.3: Standard Error of Judgment for Panel Cut Scores

Panel	SEM	SEM / 2	SE_j (all ratings)		
			A2/B1	B1/B2	B2/C1
Listening (US)	4.52	2.26	2.55	0.95	0.74
Listening (UK)			3.50	1.22	1.11
GCVR (US)	3.96	1.98	2.23	0.80	0.42
GCVR (UK)			2.03	1.45	3.00
Writing	5.14	2.57	0.55	0.70	0.00
Speaking	0.82	0.41	0.00	0.00	0.08

The table shows that the SEj values are much less than half of each section's SEM value for most of the panels' cut scores. This suggests that these cut score recommendations are dependable and that they would likely be replicated in another standard setting study. However, the table also shows that the SEj values for the A2/B1 cut scores from both the US and UK listening and GCVR panels, and the B2/C1 cut score from UK GCVR panel are slightly higher than half of each sections' SEM value. This suggests that these cut score recommendations are somewhat less likely to be replicated in another standard setting study. One possible way to reduce the SEj values would be to remove some of the more extreme cut score recommendations (those that were too high or too low compared to the rest of the panel) from the overall cut score calculations. This is typically done when one panelist's outlying cut score recommendation exerts too much influence on the overall cut score recommendation. While this option was considered, analysis of Tables 3.8, 3.9, 3.10, and 3.11 reveal that all of the panelists' cut score recommendations were quite similar, and that none of the estimates were really outlying enough to justify excluding them from the cut score calculations. Therefore, despite the fact that the SEj estimates for these cut scores were slightly larger than half of the SEM, the data suggests that, as a whole, the panelists' cut score recommendations are dependable and would likely be replicated in another standard setting study.

Analysis of the decision consistency can provide another piece of internal validity evidence. To measure this consistency, this report utilizes the methods and tables presented in Subkoviak (1988) to estimate the agreement coefficient (p_0) and kappa coefficient (κ) for each cut score. Both of these coefficients measure classification consistency; they just do it in slightly different ways. The agreement coefficient is a measure of overall consistency that represents the proportion of test takers that would be consistently classified on two administrations of the same test (Subkoviak, 1988). The kappa coefficient is a measure of the test's contribution to that consistency, and this gain in consistency is expressed as a percentage of maximum possible gain (Subkoviak, 1988).

The summary statistics and reliability estimates from above were also used here, in conjunction with the formula for calculating standard z scores and the tables from Subkoviak (1988), to estimate the agreement and kappa coefficients for each cut score.

Table 4.4: Agreement Coefficient (p_0) and Kappa (κ) for Panel Cut Scores

Cut Score	Listening		GCVR		Writing		Speaking	
	p_0	κ	p_0	κ	p_0	κ	p_0	κ
B2/C1	0.92	0.59	0.92	0.68	0.94	0.49	0.81	0.58
B1/B2	0.84	0.65	0.86	0.71	0.80	0.59	0.93	0.50
A2/B1	0.90	0.61	0.90	0.68	0.83	0.57	-	-

Table 4.4 summarizes these estimates for each section's overall cut scores. It should be noted that for high-stakes exams such as the MELAB, reliability estimates of 0.80 and above are expected and acceptable. Thus, based on Subkoviak's (1988) tables, we should expect agreement coefficients greater than or equal to 0.80, and kappa coefficients between 0.45 and 0.71. Table 4.4 shows that the agreement and kappa coefficients are generally quite high ($p0 \geq 0.80$, $\kappa \geq 0.49$) for each cut score, except for the speaking section A2/B1 cut score. Agreement and kappa estimates could not be obtained for this cut score because there weren't enough 2015 speaking test scores at the A2 level to make any claims about the

4 The MELAB speaking section is only scored by one examiner, so SEM estimates could not be obtained using regular test administration data.

5 The UK writing panel is not included in this analysis because the panel did not make direct judgments on the cut scores, but instead scored the essays using the CAE writing rating scale.

cut score's performance. However, the strong agreement coefficients of the other cut scores suggest that test takers would likely be consistently classified into the same CEFR level if they were to take the exam multiple times, and the strong kappa coefficients of these cut scores suggest that the reliability of the test scores has a good contribution to the overall classification consistency.

Overall, this section has provided two important pieces of internal validity evidence. The analysis of the SEj values provides evidence that the recommended cut scores are replicable, and the decision consistency analysis provides evidence that the test can consistently classify test takers with these recommended cut scores. These two pieces of internal validity evidence work to support the overall quality of the cut score recommendations.

4.3 EXTERNAL VALIDITY

This section summarizes the available external validity evidence to provide support for the recommended MELAB cut scores. This kind of validity evidence is often the most difficult to obtain (Council of Europe, 2009, Ch. 7). It typically consists of independent evidence that supports the results of the standard setting study (Council of Europe, 2009, Ch. 7), such as cut score recommendations obtained using a different standard setting method or the results from an external measure of the test takers' language ability (e.g., results from another CEFR-linked test, CEFR judgments by teachers) to compare with MELAB results. Unfortunately no external measures of the language ability were available, and applying a second standard setting method would have greatly increased the complexity of the judgment task, making it more difficult and time consuming for the panelists.

Even so, this report attempts to provide some external validity evidence by exploring the reasonableness of the recommended cut scores. This was done by applying the recommended cut scores to the 2015 MELAB test results and examining the resulting CEFR distributions. Table 4.5 presents the CEFR distributions for each section. It shows that the CEFR distributions for the required test sections (listening, GCVR, and writing) were all very similar, with most of the test takers scoring at the B1 and B2 levels. This is similar to our expectations, since the MELAB is primarily used for education program admissions. The CEFR distribution of the speaking section differs greatly from those of the other three sections, with most of the test takers scoring at the B2 and C1 levels. This, too, is similar to our expectations, because typically the test takers who opt to take the speaking test are more proficient users of English than those who opt not to. Overall, these CEFR distributions help to provide some external validity evidence for the MELAB's recommended cut scores.

Table 4.5: CEFR Distribution (in %) of 2015 MELAB Test Takers Based on the Recommended Cut Scores

Section	A2 or Below	B1	B2	C1
Listening	15.82	45.18	29.01	9.99
GCVR	18.46	39.76	28.04	13.74
Writing	18.95	44.90	32.96	3.19
Speaking*	0.21	10.42	47.24	42.13

*Note that the speaking test is an optional component

5. CONCLUSION

This report has provided a detailed summary of the multi-panel standard setting study conducted to link MELAB test scores to the CEFR. It has documented both the procedures and results of the study, including the standard setting meetings, and has provided procedural, internal, and external validity evidence to support the quality of the cut score recommendations. The raw cut scores are finalized by converting the raw listening and GCVR cut scores to MELAB scaled scores and by transforming the writing and speaking cut scores to the corresponding values on the MELAB writing and speaking rating scales. Table 5.1 presents the CEFR score bands, based on these cut score recommendations, for each section of the MELAB.

Table 5.1: Final MELAB CEFR Score Bands

CEFR Level	Listening	GCVR	Writing	Speaking*
C1	92 – 100	91 – 100	90 – 97	4- – 4
B2	81 – 91	78 – 90	77 – 87	3- – 3+
B1	62 – 80	57 – 77	67 – 75	2- – 2+
A2 or Below	0 – 61	0 – 56	0 – 65	1 – 1+

*Note that the speaking scale includes "plus" and "minus" bands between 1 and 4.
The full scale is 1, 1+, 2-, 2, 2+, 3-, 3, 3+, 4-, 4

6. REFERENCES

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting Performance Standards: Contemporary Methods. *Educational Measurement: Issues and Practice*, Winter 2004, 31–50.

Cizek, G. J. & Bunch, M. B. (2007). *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*. Thousand Oaks, CA: Sage.

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

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

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

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge: Press Syndicate of the University of Cambridge.

De Jong, J. H. A. L. (2013, May). *Extending and complementing the Common European Framework*. Paper presented at the European Association for Language Testing and Assessment, Istanbul.

Kaftandjieva, F. (2010). *Methods for Setting Cut Scores in Criterion-referenced Achievement Tests: A comparative analysis of six recent methods with an application to tests of reading in EFL*. Arnhem: Cito.

Lim, G. S., Geranpayeh, A., Khalifa, H., & Buckendahl, C. W. (2013). Standard setting to an international reference framework: Implications for theory and practice. *International Journal of Testing*, 13(1), 32–49.

Mills, C. N., Melican, G. J., & Ahulwalia, N. T., (1991). Defining Minimal Competence. *Educational Measurement: Issues and Practice*, 10(2), 7–10, 14.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 249–281). Mahwah, NJ: Erlbaum.

Morrow, K. (2004). Background to the CEF, in Morrow, K. (Ed.). I*nsights from the Common European Framework* (pp 3–11), Oxford: Oxford University Press.

North, B. (2014). *The CEFR in Practice*. Cambridge: Cambridge University Press.

Papageorgiou, S. (2010). Investigating the decision-making process of standard setting participants. *Language Testing* 27(2), 261–282.

Papageorgiou, S., Tannenbaum, R., Bridgeman, B., & Cho, Y (2015). The association between TOEFL iBT test scores and the Common European Framework of Reference (CEFR) levels. (Research Memorandum No. RM-15-06). Princeton, NJ: ETS.

Skorupski, W. P. (2012). Understanding the cognitive processes of standard setting panelists (p. 135-147). In G. J. Cizek (Ed.) *Setting Performance Standards: Foundations, Methods, and Innovations*. Routledge: NY, NY.

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

Tannenbaum, R. J. (2011). Standard setting. In J. W. Collins & N. P. O'Brien (Eds.), *Greenwood dictionary of education* (2nd ed., p. 441). Santa Barbara, CA: ABC-CLIO.

Tannenbaum, R. J. & Cho, Y. (2014). Critical factors to consider in evaluating standard-setting studies to map language test scores to frameworks of language proficiency. *Language Assessment Quarterly,* 11(3), 233–249.

Tannenbaum R. J. & Wylie E. C. (2008). *Linking English-language test scores onto the Common European Framework of Reference: An application of standard-setting methodology*. ETS RM-08-34, Princeton, NJ: Educational Testing Service. Retrieved from: https://www.ets.org/Media/Research/pdf/RR-08-34.pdf

APPENDIX A: CEFR SCALES USED FOR EACH MELAB SKILL PANEL

Table A.1: CEFR Scales Used in US Listening Section Familiarization Activities

CEFR Scale	Page Number (Council of Europe, 2001)	Familiarization Activity
Overall Listening Comprehension	66	1
Understanding a Native Speaker Interlocutor	75	1
Understanding Conversation Between Native Speakers	66	2
Listening as a Member of a Live Audience	67	2
Listening to Announcements and Instructions	67	2
Listening to Audio Media and Recordings	68	2
Watching TV and Film	71	2
Formal Discussion and Meetings	78	2
Goal-Oriented Co-operation	79	2

Table A.2: CEFR Scales Used in US GCVR Section Familiarization Activities

CEFR Scale	Page Number (Council of Europe, 2001)	Familiarization Activity
Overall Reading Comprehension	69	1
General Linguistic Range	110	1
Reading for Orientation	70	2
Reading for Information and Argument	70	2
Identifying Cues and Inferring (Spoken & Written)	72	2
Grammatical Accuracy	114	2
Vocabulary Range	112	2
Vocabulary Control	112	2

Table A.3: CEFR Scales Used in US Writing Section Familiarization Activities

CEFR Scale	Page Number (Council of Europe, 2001)	Familiarization Activity
Overall Written Production	61	1
Reports and Essays	62	1
Creative Writing	62	2
Overall Written Interaction	83	2
Correspondence	83	2
General Linguistic Range	110	2
Grammatical Accuracy	114	2
Thematic Development	125	2
Coherence and Cohesion	125	2

Table A.4: CEFR Scales Used in US Speaking Section Familiarization Activities

CEFR Scale	Page Number (Council of Europe, 2001)	Familiarization Activity
Overall Oral Production	58	1
Overall Spoken Interaction	74	1
Sustained Monologue: Describing Experience	59	2
Conversation	76	2
Vocabulary Range	112	2
Vocabulary Control	112	2
Spoken Fluency	129	2

Table A.5: CEFR Scales Used in UK Listening Panel Familiarization Activities

CEFR Scale	Page Number (Council of Europe, 2001)	Familiarization Activity
Overall Listening Comprehension	66	1
Understanding a Native Speaker Interlocutor	75	1
Understanding Conversation Between Native Speakers	66	1
Listening as a Member of a Live Audience	67	2
Listening to Announcements and Instructions	67	2
Listening to Audio Media and Recordings	68	2

Table A.6: CEFR Scales Used in UK GCVR Panel Familiarization Activities

CEFR Scale	Page Number (Council of Europe, 2001)	Familiarization Activity
Overall Reading Comprehension	69	1
General Linguistic Range	110	1
Reading for Orientation	70	2
Reading for Information and Argument	70	2
Grammatical Accuracy	114	3
Vocabulary Range	112	3
Vocabulary Control	112	3

APPENDIX B: EXAMPLE PRE-STUDY ACTIVITY

Approximately five days prior to each study meeting, panel participants were sent this activity (in conjunction with two CEFR scales) to complete before the meeting convened:

1. Based on the information in the CEFR Global Scale and Self-Assessment Grid[1], please describe what you perceive are the key characteristics of an **average B1 writer**[2].

_____

2. Based on the information in the CEFR Global Scale and Self-Assessment Grid, please describe what you perceive are the key characteristics of a **just-qualified (i.e., minimally competent) B1 writer**.

_____
_____

3. Based on the information in the CEFR Global Scale and Self-Assessment Grid, please describe what you perceive are the key characteristics of an **average B2 writer**.

_____
_____

4. Based on the information in the CEFR Global Scale and Self-Assessment Grid, please describe what you perceive are the key characteristics of a **just-qualified (i.e., minimally competent) B2 writer**.

_____
_____

5. Based on the information in the CEFR Global Scale and Self-Assessment Grid, please describe what you perceive are the key characteristics of an **average C1 writer**.

_____
_____

6. Based on the information in the CEFR Global Scale and Self-Assessment Grid, please describe what you perceive are the key characteristics of a **just-qualified (i.e., minimally competent) C1 writer**.

_____
_____

1 Note that for the speaking panel, the panelists were asked to refer to the CEFR Global Scale and the Qualitative Aspects of Spoken Language Use table.

2 For the speaking panel, the term "speaker" was used in all six questions, for the listening panel the term "listener" was used, and for the GCVR panel, the term "reader" was used.

APPENDIX C: FAMILIARIZATION ACTIVITY RESULTS

Table C.1: US Listening Panel Familiarization Activity 1 Results (21 Descriptors, 3.33 Average CEFR Level)

Measure	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13	Average
# Correct	12	16	14	13	11	12	17	11	15	14	12	15	14	13.5
% Correct	57.1	76.2	66.7	61.9	52.4	57.1	81.0	52.4	71.4	66.7	57.1	71.4	66.7	64.5
Correlation (ρ)	0.86	0.95	0.95	0.92	0.87	0.92	0.97	0.86	0.95	0.90	0.90	0.92	0.93	0.91
Average	3.57	3.29	3.57	3.67	3.86	3.57	3.52	3.62	3.52	3.43	3.62	3.48	3.62	3.56

Table C.2: US Listening Panel Familiarization Activity 2 Results (40 Descriptors, 3.38 Average CEFR Level)

Measure	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13	Average
# Correct	24	18	23	19	19	23	23	19	24	20	24	17	26	21.5
% Correct	60.0	45.0	57.5	47.5	47.5	57.5	57.5	47.5	60.0	50.0	60.0	42.5	65.0	53.7
Correlation (ρ)	0.91	0.85	0.80	0.86	0.81	0.89	0.92	0.81	0.87	0.86	0.89	0.77	0.87	0.85
Average	3.55	3.5	3.65	3.4	3.9	3.78	3.25	3.83	3.68	3.65	3.38	3.73	3.63	3.61

Table C.3: US GCVR Panel Familiarization Activity 1 Results (20 Descriptors, 3.35 Average CEFR Level)

Measure	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	Average
# Correct	16	18	9	13	16	13	16	16	13	17	16	9	18	14.6
% Correct	80.0	90.0	45.0	65.0	80.0	65.0	80.0	80.0	65.0	85.0	80.0	45.0	90.0	73.1
Correlation (ρ)	0.97	0.99	0.87	0.84	0.96	0.91	0.97	0.96	0.92	0.97	0.93	0.90	0.99	0.94
Average	3.45	3.25	3.6	3.35	3.45	3.25	3.25	3.45	3.65	3.3	3.3	3.2	3.25	3.37

Table C.4: US GCVR Panel Familiarization Activity 2 Results (46 Descriptors, 3.30 Average CEFR Level)

Measure	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	Average
# Correct	25	29	26	23	36	23	27	26	24	32	33	36	33	28.7
% Correct	54.4	63.0	56.5	50.0	78.3	50.0	58.7	56.5	52.2	69.6	71.7	78.3	71.7	62.4
Correlation (ρ)	0.89	0.93	0.86	0.89	0.96	0.88	0.94	0.90	0.91	0.93	0.94	0.94	0.95	0.91
Average	3.3	3.46	3.37	3.28	3.5	3.2	3.07	3.22	3.15	3.33	3.2	3.24	3.41	3.29

Table C.5: US Writing Panel Familiarization Activity 1 Results (17 Descriptors, 4.00 Average CEFR Level)

Measure	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14	Average
# Correct	12	9	12	9	9	10	13	10	11	8	10	12	7	15	10.5
% Correct	70.6	52.9	70.6	52.9	52.9	58.8	76.5	58.8	64.7	47.1	58.8	70.6	41.2	88.2	61.8
Correlation (ρ)	0.91	0.75	0.67	0.76	0.79	0.78	0.92	0.66	0.73	0.66	0.89	0.87	0.87	0.97	0.80
Average	3.82	3.94	3.76	3.76	3.82	3.88	3.71	3.65	3.65	3.76	3.65	3.76	4.12	3.88	3.80

Table C.6: US Writing Panel Familiarization Activity 2 Results (58 Descriptors, 3.26 Average CEFR Level)

Measure	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14	Average
# Correct	31	24	30	28	24	37	28	30	32	34	31	34	35	30	30.6
% Correct	53.5	41.4	51.7	48.3	41.4	63.8	48.3	51.7	55.2	58.6	53.5	58.6	60.3	51.7	52.7
Correlation (ρ)	0.89	0.77	0.88	0.84	0.80	0.91	0.88	0.90	0.88	0.89	0.88	0.87	0.89	0.89	0.87
Average	3.29	3.52	3.53	3.09	3.62	3.45	3.17	3.62	3.47	3.34	3.22	3.45	3.28	3.43	3.39

Table C.7: US Speaking Panel Familiarization Activity 1 Results (28 Descriptors, 3.54 Average CEFR Level)

Measure	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	Average
# Correct	20	18	21	19	21	18	18	18	16	22	18	16	20	18.8
% Correct	71.4	64.3	75.0	67.9	75.0	64.3	64.3	64.3	57.1	78.6	64.3	57.1	71.4	67.3
Correlation (ρ)	0.96	0.91	0.93	0.94	0.95	0.90	0.95	0.94	0.91	0.94	0.92	0.90	0.96	0.93
Average	3.75	3.79	3.61	3.71	3.64	3.75	3.75	3.46	3.57	3.36	3.54	3.79	3.39	3.62

Table C.8: US Speaking Panel Familiarization Activity 2 Results (59 Descriptors, 3.15 Average CEFR Level)

Measure	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	Average
# Correct	41	43	35	29	35	38	39	35	31	27	41	35	40	36.1
% Correct	69.5	72.9	59.3	49.2	59.3	64.4	66.1	59.3	52.5	45.8	69.5	59.3	67.8	61.1
Correlation (ρ)	0.92	0.95	0.92	0.91	0.85	0.93	0.90	0.92	0.88	0.77	0.92	0.90	0.92	0.90
Average	3.1	3.29	2.92	3.42	2.95	3.39	3.17	2.88	3.37	3.44	3.00	3.19	3.15	3.17

Table C.9: UK Listening Panel Familiarization Activity 1 Results (26 Descriptors, 3.29 Average CEFR Level)

Measure	L1	L2	L3	L4	L5	Average
# Correct	20	20	19	13	15	17.4
% Correct	76.9	76.9	73.1	50.0	57.7	66.9
Correlation (ρ)	0.95	0.99	0.96	0.92	0.93	0.95
Average	3.33	3.33	3.38	3.38	3.38	3.36

Table C.10: UK Listening Panel Familiarization Activity 2 Results (19 Descriptors, 3.53 Average CEFR Level)

Measure	L1	L2	L3	L4	L5	Average
# Correct	14	19	17	16	16	16.4
% Correct	73.7	100.0	89.5	84.2	84.2	86.3
Correlation (ρ)	0.91	1.00	0.97	0.96	0.96	0.96
Average	3.37	3.53	3.42	3.47	3.47	3.45

Table C.11: UK GCVR Panel Familiarization Activity 1 Results (20 Descriptors, 3.35 Average CEFR Level)

Measure	L1	L2	L3	Average
# Correct	17	15	17	16.3
% Correct	85.0	75.0	85.0	81.7
Correlation (ρ)	0.98	0.96	0.99	0.98
Average	3.40	3.40	3.20	3.3

Table C.12: UK GCVR Panel Familiarization Activity 2 Results (17 Descriptors, 2.94 Average CEFR Level)

Measure	L1	L2	L3	Average
# Correct	14	14	11	13.0
% Correct	82.4	82.4	64.7	76.5
Correlation (ρ)	0.96	0.94	0.89	0.93
Average	3.00	2.88	3.06	3.0

Table C.13: UK GCVR Panel Familiarization Activity 3 Results (24 Descriptors, 3.54 Average CEFR Level)

Measure	L1	L2	L3	Average
# Correct	19	22	15	18.7
% Correct	79.2	91.7	62.5	77.8
Correlation (ρ)	0.97	0.96	0.95	0.96
Average	3.58	3.42	3.25	3.4