Linking the Common European Framework of Reference and the MET Writing Test

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

11/2014

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1 Overview

This report presents the results of a standard setting study conducted in October of 2014. The purpose of this study was to link the Michigan English Test's (MET) writing section scores to the proficiency levels of the Common European Framework of Reference (CEFR; Council of Europe, 2001). This study followed the procedures and guidelines outlined in the manual prepared by the Council of Europe to support standard-setting (Council of Europe, 2009). It establishes minimum MET writing cut scores for the appropriate CEFR proficiency levels. This report documents how the study was conducted and provides evidence of the quality and validity of the study.

1.2 Common European Framework

The Common European Framework of Reference (CEFR; Council of Europe, 2001) provides a common basis for evaluating the achievement and ability level of foreign language learners. It 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: 1). Developed during a large research project that spanned several years (North, 2000; North & Schneider, 1998), the CEFR describes six main levels of proficiency: A1 (the lowest) to C2 (the highest). Since its introduction, the CEFR has become widely used to interpret test scores. Test users and other stakeholders find the CEFR levels useful for decision-making. Linking the MET writing test to the CEFR will help test users to better interpret the test results.

1.3 Standard Setting

Standard setting is defined as the decision-making process of classifying candidates into a number of levels or categories (Kane, 2001: 53). The "boundary between adjacent performance categories" (Kane et al., 1999: 344) is called a cut score, and can be described as the "point on a test's score scale used to determine whether a particular score is sufficient for some purpose" (Zieky et al., 2008: 1). For example, when determining whether candidates have passed or failed an exam, a cut score functions as the boundary between the pass and fail category.

During a standard-setting meeting, a panel of expert judges (often called panelists) makes judgments on which examination providers will base their final cut score decisions. Under the guidance of one or more meeting facilitators, the panelists go through the process of determining the cut scores. The first stage, familiarization, requires the panelists to learn and understand both the exam and the CEFR. The second stage, training, provides panelists with the opportunity to practice and understand the procedure that will be used for making judgments. The final stage, judgment, is where the panelists make decisions on the cut scores. More than one round of judgments is organized to allow panelists to discuss their decisions and, if necessary, make adjustments. The result of the standard-setting meeting is the recommended cut scores that link the exam to the CEFR.

Once the standard-setting activities are completed, the meeting is evaluated for procedural, internal, and external validity (Council of Europe, 2009: Ch. 7). Procedural validity establishes that the procedures followed were practical and implemented properly, that the feedback given to the judges was effective, and that documentation had been sufficiently compiled. Internal validity addresses issues of accuracy and consistency of the standard setting results. External validity is provided by evidence from independent sources which support the outcome of the standard setting meeting.

1.4 Michigan English Test

The Michigan English Test (MET) is a standardized, multi-level examination of general English language proficiency provided by Cambridge Michigan Language Assessments (CaMLA). The MET listening and reading sections measure listening, reading, grammar, and vocabulary skills in personal, public, occupational, and educational contexts. It is intended for adults and adolescents at or above a secondary level of education who want to measure their general English language proficiency in a variety of linguistic contexts. The exam covers a range of proficiency levels, from upper beginner to lower advanced; the A2 to C1 levels of the CEFR (Papageorgiou, 2010).

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. CaMLA works closely with test centers to ensure that its tests are administered in a way that is fair and accessible to candidates and that the MET is open to all people who wish to take the exam, regardless of the school they attend.

1.5 MET Writing Test

The focus of this study is on an addition to the MET suite: the MET writing test. Like the MET speaking test, this section is an optional component of the MET. The purpose of the MET writing test is to evaluate a test taker's ability to write in English. It is designed to measure the writing proficiency of English language learners from upper beginner to lower advanced (A2 to C1). In order to measure the writing proficiency of individuals at these differing levels of ability, the MET writing test requires test takers to produce written language at the sentence, paragraph, and essay levels. The test taker's performance is evaluated by a certified rater, and the final score reflects the test taker's overall writing proficiency in English.

The MET writing test takes 45 minutes to complete and consists of two independent tasks, each scored analytically on five different criteria (grammatical accuracy, vocabulary, mechanics, organization and connection of ideas, and task completion). In Task 1, the test taker is presented with three questions on a connected theme. These three questions require the test taker to respond with a series of sentences that connect ideas together. This task is accessible to lower proficiency writers who can produce written text at the sentence level, but who may struggle to connect ideas together beyond a simple paragraph. In Task 2, the test taker is presented with a single writing prompt designed to elicit a short essay from the test takers. This task offers more proficient writers the scope to show the full range of their abilities. It evaluates the test taker's ability to compose an essay that consists of several paragraphs. Both tasks are scored using the same analytic rating scale, and the results are added together to obtain the test taker's final score, which ranges from 0 to 40.

2. METHODOLOGY

2.1 Panelists

A key feature of this study is that it employed two independent panels of judges. This approach has been used in previous standard setting studies (e.g. Tannenbaum & Katz, 2008; Brunfaut & Harding, 2014), and was selected because it allowed for the comparison and cross-validation of the two panels' recommended cut scores. In total, 17 panelists participated in this study, 9 in one panel, 8 in the other. The panels provided an experienced and diverse representation of professionals in the field of English as a second language and represented a range of stakeholders (curriculum directors, testing coordinators, teachers, and assessment specialists).

The first panel (henceforth referred to as Panel 1) comprised 9 participants, selected from the Centro Colombo Americanos in Colombia.[1] The panelists were all experienced teachers of English as a second language (ranging from 9–31 years), with undergraduate and graduate qualifications in education, linguistics, and other related fields. They reported a range of language testing experience, primarily as test administrators or interviewers, but several panelists also indicated that they had experience as writing examiners.

The second panel (henceforth referred to as Panel 2) consisted of 8 participants, selected from within CaMLA, that are experts in the field of language assessment. Like Panel 1, they have undergraduate and graduate qualifications in education, linguistics, and applied linguistics, and have a range of experience as language teachers (1–10 years). Many of these panelists have experience as writing examiners or raters on other English language testing exams.

For both panels, the experience and understanding of the CEFR levels was varied, so it was important to ensure the panelists were properly calibrated though familiarization tasks.

2.2 Standard-Setting Method

The standard-setting method used in this study is a modification of the bookmark method. Developed in 1996, the bookmark method is an item response theory based standard setting procedure that has recently gained popularity in the United States (Lewis et al., 2012). One of its key feature is that it is a test centered method that is applicable not only to multiple choice items, but also constructed response items (Council of Europe, 2009: Ch. 6). A key component of this method is the ordered item booklet (OIB), which contains several items listed in order of increasing difficulty (typically using the IRT difficulty parameter estimate). Panelists make judgments by reading through the OIB and placing bookmarks at the first item that a borderline or just qualified candidate would be "unlikely" to answer. In the context of the bookmark method, "unlikely" means that the probability of a

1 12 panelists participated in different stages of the meeting but a complete dataset was available for only nine participants.

borderline candidate correctly answering the question is less than the predefined mastery criterion, called the response probability (RP; often set at 2/3). Panelists place their bookmarks at the item in the OIB where they feel a borderline candidate's probability of successfully obtaining the correct response is less than the response probability. Because the items are placed in order of increasing difficulty, the borderline candidate would be expected to correctly answer the questions placed up to the bookmark.

This study applied the bookmark method to the two writing tasks on the MET writing test. Each task had its own OIB. Separate sets of cut scores (A2/B1, B1/B2, and B2/C1) were determined for each task/for Task 1 and Task 2. The cut scores for each task were added together to obtain the cut scores for the exam. The OIBs consisted of responses to the task for score points ranging from 5 to 20. The bookmark method was selected for this study because it is a straightforward approach to obtaining cut scores for constructed response items. Additionally, this method allows the panelists to make judgments on multiple cut scores without making unreasonable cognitive demands.

2.3 Standard-Setting Procedure

This section outlines the procedures conducted during this standard setting study before, during, and after the standard setting meetings. Though the two panel approach used by this study necessitated two separate standard setting meetings, they were conducted one week apart and were presided over by the same meeting facilitator. The goal was to ensure that the two meetings were conducted in the same way.

Prior to the standard setting meeting, the panelists were sent a packet of materials to familiarize themselves with the MET writing test and the CEFR. These materials included a sample MET writing test, the MET writing test rating scale, a table of CEFR level descriptors (Council of Europe, 2001: 26–27), and an article on the background to the CEFR (Morrow, 2004). Additionally, the panelists were asked to think about students at the beginning of CEFR levels A2, B1, B2, and C1 and answer the following questions:

- What should you expect students at the beginning of these levels to be able to do if writing in English?
- What in-class behaviors would you observe to let you know the level of the student's writing proficiency?
- What characteristics define students with "just enough" English writing skills to enter into each of these three CEFR levels?

The panelists were asked to review these materials and questions prior to the standard setting meeting so that any questions or comments could be discussed. The panelists were also given a background questionnaire that was to be completed and submitted to the researchers before the meeting.

The standard setting meeting began with an introduction to the study's goals, the CEFR, the MET writing test, and the linking process, as well as a discussion of the prereading materials. Once this was completed, the panelists received the familiarization task. For this task the panelists received "atomized" CEFR descriptors, which they independently assigned to the appropriate CEFR level (A2–C2). The "atomization" of the descriptors into short statements is based on Kaftandjieva and Takala (2002); it aims to familiarize the panelists with all constituent statements of the descriptors. This task is challenging for the panelists because several descriptor statements are quite short and do not contain a detailed description of the context of language use. However, it encourages careful reading of (and therefore familiarization with) the details of the CEFR writing descriptors.

Once the panelists had assigned CEFR levels to the descriptors, the results were condensed into a single excel file. The panelists were then shown how many descriptors they placed at the correct level and how their mean CEFR level compared with the true mean level. The meeting facilitator encouraged the participants to discuss their answers and explain their reasons for choosing a particular level, particularly for descriptors where there was a high amount of disagreement with the correct CEFR level. The discussion went through all of the descriptors, and only moved on to the next one when all of the panelists felt that they understood the correct CEFR level of a descriptor statement. The panelists were given a handout with the correct CEFR levels as a set of Performance Level Descriptions (PLD; see Cizek & Bunch, 2007: 44–47). These would be used in the subsequent tasks as guidance when panelists made their cut score decisions. To further familiarize themselves with the CEFR, the panelists were asked

to define the just qualified candidate at the A2, B1, B2, and C1 levels. Using the CEFR descriptors as a guide, the panelists collaboratively developed a list that described what they expected a just qualified candidate would be able to do at each CEFR level.

The next step, the training task, provided panelists with the opportunity to practice making judgments with the bookmark method. Each panelist received two short ordered item booklets (OIB), one for each of the two writing tasks, which contained a selection of responses with scores ranging from 9 to 17. Using these OIBs, the panelists practiced making their cut score judgments for candidates at the A2/B1 and B1/B2 borders, first for Task 1 of the MET writing test, and then for Task 2.The results for all panelists were summarized in an excel file, and the cut scores for the two tasks were combined into a total cut score. The panelists discussed their results and any questions about the judgment procedure were addressed. At the conclusion of the training task the panelists were given a pre-judgment questionnaire to assess their understanding of the standard setting procedure and their willingness to proceed to the next task.

The final stage of the standard setting meeting was the judgment task, where the panelists determined the recommended cut scores for linking the MET writing test to the CEFR. The procedure for making the judgments was the same as in the training task. For the judgment phase the panelists were required to make judgments for candidates at the A2/B1, B1/B2, and B2/C1 borders. The panelists received a complete OIB for each task (Task 1 and Task 2). Each OIB contained performances with scores from 5 to 20 on the rating scale. There were two rounds of judgments. After each round, the panelists discussed the results and shared their rationale. They also reviewed the CEFR distribution of the pilot data based on their recommended overall cut scores. The end result of the judgment task was three recommended cut scores (A2/B1, B1/B2, B2/C1) that were obtained by adding together the cut scores for Task 1 and Task 2. The standard setting meeting concluded with a post-judgment questionnaire in which panelists gave their opinions on the overall quality of the meeting, the usefulness of the methods employed, and their confidence in the final recommended cut scores.

After the standard setting meetings, the results of the familiarization and judgment tasks were analyzed, along with the pre-judgment and post-judgment questionnaires. The final recommended cut scores from both panels were compared. The two independent panels had recommended similar cut scores. Consequently, the final MET writing test cut scores were obtained by taking the average of the two panels recommended cut scores. Evidence of the procedural, internal, and external validity of the cut scores was also obtained and summarized.

3. RESULTS

3.1 Specification

In addition to presenting the results of the standard setting meeting, it is important to provide "evidence that the types of language skills measure by the test are consistent with those described by the framework" (Tannenbaum and Cho, 2014: 237). This stage is referred to as specification by the Council of Europe's manual for relating exams to the CEFR (2009: Ch. 4), and as construct congruence by Tannenbaum and Cho (2014). The specification stage requires the test designer to provide evidence that the content of the exam is in fact aligned with the CEFR. That is, there needs to be justification for the linking of the exam and the framework.

The MET writing test can be linked to the CEFR because the exam was specifically developed to assess test taker's ability to communicate in written English at levels equivalent to those described as the A2–C1 levels on the CEFR. The CEFR has been used by the development team as a reference against which to define the writing construct reflected in the MET, throughout the test's development. The test has been designed to elicit the functions that are typical of the A2–C1 CEFR levels, particularly those linguistic functions that distinguish one level from another. Because the MET writing test targets a wide range of CEFR proficiency levels two tasks were designed which, when used together, could elicit writing from low-proficiency to high-proficiency writers. Task 1 of MET writing test consists of three parts that require test takers to respond with a series of sentences that connect ideas together and Task 2 is a writing prompt designed to elicit a short essay. Table 3.1 provides a summary of the specific CEFR functions being targeted by each task. A sample prompt is available on the CaMLA website.

As shown in Table 3.1, Parts 1 and 2 of Task 1 require the test taker to describe multiple aspects of a situation or opinion and link them in a linear fashion.

Table 3.1: MET Writing Test Targeted CEFR Functions

Task	Part	Item Description	Targeted CEFR Function
T1	1	Describe a personal experience	• A2: "Can write a series of simple phrases and sentences linked with simple connectors like and, but, and because" (Council of Europe, 2001: 61). • A2: "Can use the most frequently occurring connectors to link simple sentences in order to tell a story or describe something as a simple list of points" (Council of Europe, 2001: 125).
	2	Express a personal opinion	
	3	Elaborate upon a fact or opinion	• B1: "Can use a variety of linking words efficiently to mark clearly the relationships between ideas" (Council of Europe, 2001: 125). • B1: "Can write straightforward connected texts on a range of familiar subjects within a field of interest, by linking a series of shorter discrete elements into a linear sequence" (Council of Europe, 2001: 61)
T2	1	Compare and contrast essay with reasons and examples	• B2: "Can develop an argument, giving reasons in support of or against a particular point of view and explaining the advantages and disadvantages of various options" (Council of Europe, 2001: 62). • B2: "Can support main points with relevant supporting detail and examples" (Council of Europe, 2001: 125). • C1: "Can expand and support points of view at some length with subsidiary points, reasons and relevant examples" (Council of Europe, 2001: 62). • C1: "Can integrate sub-themes, develop particular points and round off with an appropriate conclusion" (Council of Europe, 2001: 125).

This is within the capabilities of an A2 level test taker. The expected length of response is too brief for advanced textual organization, but does provide ample space for writers above the A2 level to demonstrate their clause-level and word-level competencies. Part 3 targets the B1 level of writing proficiency and requires more complex propositional content and linking as well as elaboration. The question allows test takers at the B2 level and above to demonstrate their ability to create topical and thematic unity by referencing the text they produced in Parts 1 and 2.

Task 2 is a compare and contrast essay. The prompt comprises a series of questions that provide support for A2 and B1 test takers so that they can attempt the task and partially complete it. However, it also provides scope for B2 and C1 test takers to demonstrate their ability to "develop an argument, giving reasons in support of or against a particular point of view and explaining the advantages and disadvantages of various options" (Council of Europe, 2001: 62).

The MET writing test's rating scale was also developed with the CEFR in mind. An analytic scoring rubric was selected to reflect the wide range of competencies that contribute to test taker's responses. The scale consists of five scoring criteria (grammatical accuracy, vocabulary, mechanics, organization and connection of ideas, and task completion) on a five point scale (0–4 per criterion, which means 0–20 points per task) that were selected to cover the major aspects of writing produced by individuals with CEFR levels ranging from A2 to C1. The scale is based upon theoretical models of L2 English writing and is designed to mirror the bands of writing ability described in the CEFR (Council of Europe, 2001: 61ff.).

3.2 Familiarization

The first step in the postmeeting analysis of the data was to establish the panelists' familiarity with the CEFR. It is important that all panelists understand the CEFR levels and can consistently rank the CEFR descriptors. If the panelists are unable to assign descriptors consistently to the correct CEFR level, then they may provide inconsistent or inaccurate judgments when setting cut scores. This would undermine the validity of the recommended cut score.

Tables 3.2 and 3.3 present information about the panelists' individual performance on the familiarization tasks for Panels 1 and 2, respectively. Similar analyses can be found in other relevant studies, such as

Table 3.2: Panel 1 Familiarization Results (47 Descriptors, 3.26 Mean CEFR Level)

	J1	J2	J3	J4	J5	J6	J7	J8	J9
Correct	21	29	27	19	22	23	18	25	25
ρ	0.76	0.90	0.83	0.84	0.80	0.85	0.83	0.91	0.86
Mean	3.77	3.45	3.51	3.57	3.74	3.74	3.66	3.43	3.51

Table 3.3: Panel 2 Familiarization Results (47 Descriptors, 3.26 Mean CEFR Level)

	J1	J2	J3	J4	J5	J6	J7	J8
Correct	33	13	24	23	29	27	22	20
ρ	0.92	0.78	0.90	0.83	0.91	0.89	0.90	0.83
Mean	3.47	3.60	3.64	3.55	3.30	3.30	3.53	3.91

Kaftandjieva and Takala (2002), Generalitat de Catalunya (2006), and Papageorgiou (2007).

The tables show the total number of correctly placed descriptors for each panelist, the Spearman correlation (ρ) between their CEFR level placement and the correct CEFR level of the descriptor, and the mean CEFR level assigned. When interpreting the information presented in these tables, it is important to note that while high correlations indicate that the panelists understand how the descriptors progress from lower to higher levels, they do not indicate how accurate the panelists were in assigning the descriptors to CEFR levels. Therefore, the correlations should be consulted in conjunction with the total number of correctly placed descriptors. The mean CEFR level assigned by each panelist shows their tendency to put descriptors at either lower or higher levels. Panelists with mean levels higher than the correct level rate more leniently while those with mean levels lower than the correct one rate more severely.

Tables 3.2 and 3.3 show that all of the correlations between the assigned CEFR levels and the descriptors' true CEFR levels are strong (> 0.75). This suggests that the panelists had a good understanding of how language proficiency progresses from lower to higher CEFR levels. However, the relatively low number of correct descriptor placements suggests that most panelists had difficulty placing the descriptors at the exact levels. All the panelists' mean CEFR levels were higher than the true mean CEFR level for the descriptors (3.26 mean CEFR level). This indicates that all of the panelists tended to be more lenient when assigning descriptors to CEFR levels. They ascribe competencies to a level higher than is actually the case.

It is important to note that all of the above findings are typical, and that they are similar to the results obtained in standard setting studies for the MET listening and reading sections (Papageorgiou, 2010) and the CaMLA EPT (CaMLA, 2014). Nevertheless, the implications of this are important for setting cut scores, since a panelist may apply their leniency (or severity) from the familiarization task to the judgment task, possibly skewing the results. This is a particular concern since all of the panelists for this study are lenient. During the discussion of the descriptor statements, the leniency of the panelists was pointed out in order to avoid skewing the judgment results.

Analysis of the familiarization task has thus far focused on the individual panelists' understanding of the CEFR levels. Since cut scores are based on the judgments from the entire panel, further analysis was performed to establish the consistency of the two panels. Table 3.4 summarizes these results.

Table 3.4: Panel Agreement and Consistency

	Panel 1	Panel 2
Alpha	0.973	0.979
ICC	0.973	0.977
W	0.800	0.846

The table presents the three most commonly used measures of agreement and internal consistency (Kaftandjieva, 2010: 96) for both panels. The first, Cronbach's Alpha, is an internal consistency index that measures "how well a group of items together measure

the trait of interest" (Davies et al., 1999: 39). In the context of standard setting, it indicates "the consistency of the reliability of ratings in terms of rater consistency" (Generalitat de Catalunya, 2006: 62). Another measure, the intraclass correlation coefficient (ICC), is also used to compare rater scores. Like Cronbach's Alpha, the ICC "takes account of the variance within and between raters in terms of harshness" (Davies et al., 1999: 89). For this study, the ICC two-way mixed model was used and average measures for exact agreement are reported. The third measure of agreement, Kendall's W, is a nonparametric statistical procedure that is often used for "the calculation of levels of agreement in situations where more than two raters are ranking the same group of subjects or attributes" (Davies et al., 1999: 100). These coefficients range from 0 to 1, with 1 indicating complete agreement and 0 indicating complete disagreement among panelists. Table 3.4 shows that these measures are high for both panels. This suggests that there is a good level of agreement and consistency among each group of panelists on the familiarization task.

The familiarization task results were also analyzed in the program, Facets, to provide further information about the consistency of the panelists in assigning the descriptors to CEFR levels (Papageorgiou, 2007). Facets is a program that "is designed to handle the more complex applications of unidimensional Rasch measurement and performs many-facet Rasch measurement" (Bond & Fox, 2007: 302). Both panels'

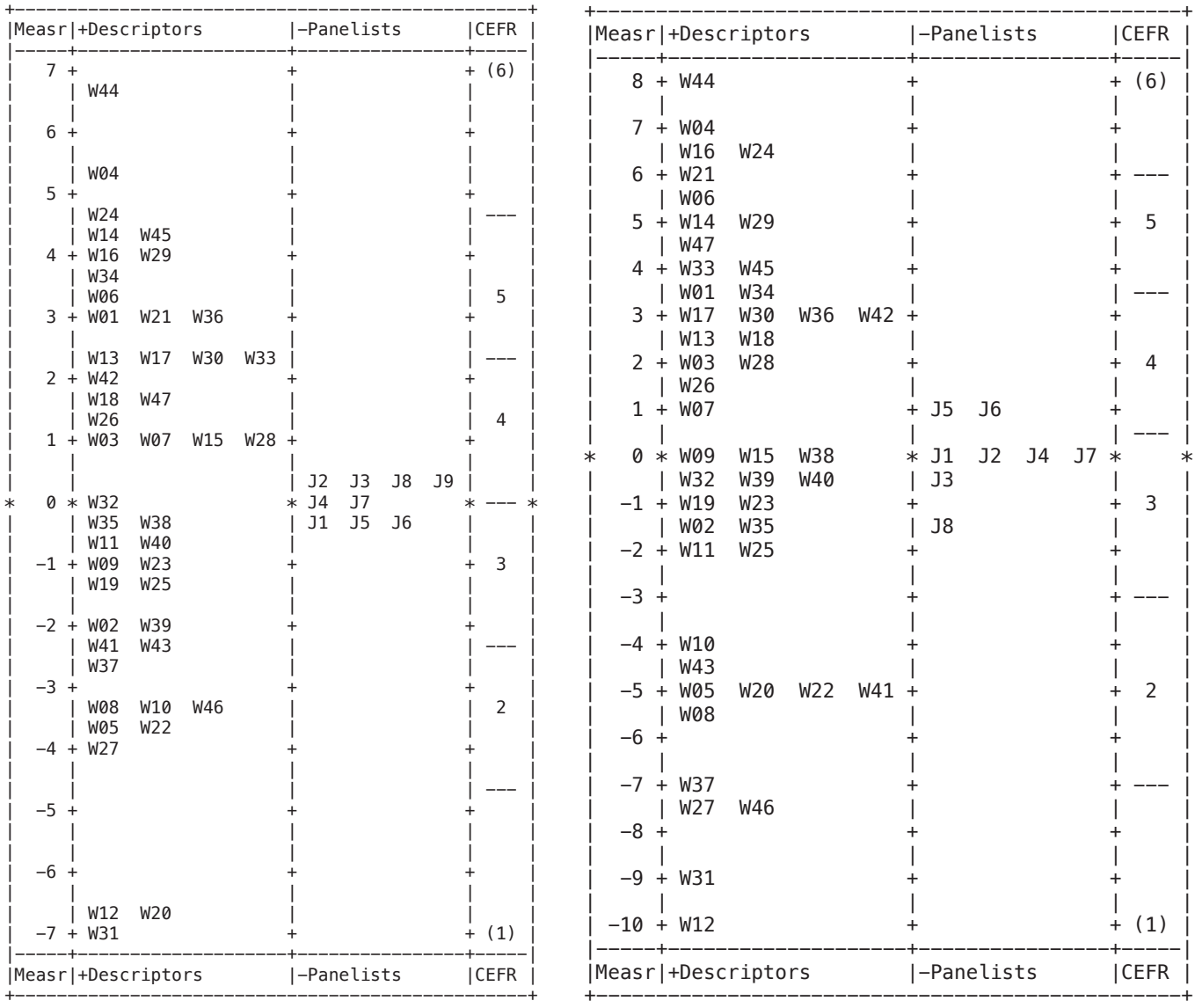

Figure 3.1: Vertical Rulers for Panel 1 (left) and Panel 2 (right)

responses to the familiarization task were analyzed in Facets. Figure 3.1 provides a graphical representation of the data in the form of the vertical ruler plots created by Facets.

When analyzing this figure it is important to understand what the columns in each plot represent. The first column (Measure) provides a common scale on which all of the facets are recorded; the Rasch measure of difficulty in logits. The second column displays the descriptor ID's according to their difficulty. More difficult descriptors (those that the judges assigned higher CEFR levels) have a larger measure, while easier ones have smaller measures. The third column lists the panelists according to their severity. More severe panelists (those who tended to assign lower CEFR levels than the others) have higher measures, while more lenient ones have lower measures. The final column presents bands for each of the CEFR levels used in the familiarization task. Analysis of Figure 3.1 reveals that while the judges on both panels were similar to each other (within a few logits), the severity/leniency of Panel 2 was more varied. Additionally, the comparatively longer Measure scale of Panel 2 suggests that their CEFR ratings were more varied than those of Panel 1. Analysis of the descriptor difficulty measures also reveals that the panels did not place the descriptors in the same order relative to one another (e.g. descriptor W20).

Table 3.5 presents information that allows for a more detailed analysis of the panelists' performance. The table lists the measure range, infit mean square range, and panelist reliability. While there are several fit statistics reported by Facets, Papageorgiou (2007: 22) states that infit is "the most meaningful" for analysis of the panelists' performance on the familiarization task. Infit is a fit statistic, sensitive to items targeted on the person, that indicates how accurately the data fits the model (Linacre, 2002). Infit mean square values are expected to be near 1. Values smaller than 1 indicate that the data has less variation and is more predictable than expected, while values larger than 1 indicate that the data has more variation and is less predictable than expected (Linacre, 2002; Papageorgiou, 2007). According to Linacre (2002), infit mean square values between 0.50 and 1.50 are considered to be productive for measurement. The reliability index here is the separation reliability. It shows the "reliable difference in severity of the raters" (Papageorgiou, 2007: 22) and differs from traditional interrater reliability in that low reliabilities are preferred when evaluating the judges.

Table 3.5: Facets Panelist Information

Panel	Measure Range	Infit Range	Reliability
Panel 1	-0.42–0.43	0.63–1.46	0.46
Panel 2	-1.33–0.85	0.59–1.65	0.82

Table 3.5 shows that the majority of the panelists had infit mean square errors within the desired range, and only the upper range of Panel 2 exceeded the upper limit of 1.50 suggested by Linacre (2002). A closer inspection of the Facets output for each panelist revealed that only one panelist (J2 from Panel 2) had an infit value that that exceeded the threshold. The excess variance of J2 is partially explained by the fact that this panelist had the most difficulty assigning descriptors to the correct CEFR level (see Table 3.3). Analysis of the measure ranges and the reliability values in Table 3.5 also reveals that Panel 2 was more varied in severity/leniency than Panel 1. This variability between panels, in addition to that found in Figure 3.1, implies that the two panels were not assigning descriptors to CEFR levels in the same way.

Additionally, as evidenced by the low numbers correct in Tables 3.2 and 3.3, individual panelists had difficulty placing descriptors at the correct level. Placing descriptors at adjacent levels is occasionally expected and understandable. Frequent errors that are the result of a systematic misunderstanding of the differences between levels raises concerns about the validity of cut scores suggested by such panelists. Nevertheless, the analysis presented in this section suggests that most panelists had a good overall understanding of the CEFR descriptors and were consistent in their use of the CEFR, particularly in how language ability progresses from lower to higher levels in the CEFR scales.

It is also important to note that the purpose of the familiarization task was to orient the panelists to the CEFR and to help them to understand the correct placement of descriptors at each CEFR level. Therefore, variation between panels at this stage is to be expected. Issues of accuracy and leniency were pointed out during the discussion, and the descriptor statements were discussed thoroughly to ensure that the panelists all understood the correct CEFR levels. Though the panels had different starting points, and demonstrated different understandings of the CEFR at the beginning of the meeting, the remainder of this report shows that the familiarization activities achieved their purpose. Based on the low variability of the judgment task (see

Section 3.3), the results of the questionnaires, and comments made throughout the discussion of the familiarization task, the process of familiarization was successful in calibrating the panelists, helping them to understand the CEFR levels, and clarifying the differences between adjacent levels.

3.3 Judgment

This section focuses on the analysis of the cut scores obtained during the judgment task. Tables 3.6 and 3.7 present the total cut scores recommended by each panelist, from Panels 1 and 2, respectively. It should be noted that these cut scores are expressed as total scores on the MET writing test, and therefore, the scores can range from 0 to 40. While the panelists made cut score judgments on both Task 1 and Task 2 of the exam, this analysis focuses only on the cut scores of interest, the resulting total cut scores.

Inspecting the mean cuts scores, we can see that the cut score recommendations from both panels are very similar. This cross-panel consistency provides validity evidence for the study and supports both panels' recommended cut scores. The standard deviations (and standard errors of judgment) were relatively small for both panels. This suggests that while each panelist recommended a different cut score, there was little variation between the panelists' recommendations. Additionally the variability in judgments (SD and SE_j) decreased after the first round of judgments (a pattern also noted by Tannenbaum and Katz, 2008). This decrease is a result of the discussion among panelists between rounds.

The end result of both standard setting meetings was a set of recommended cut scores for the MET writing test. These raw cut scores are presented in Table 3.8 for both panels, along with the overall recommended cut scores. These overall cut scores were obtained by taking the average of the two panels' recommended cut scores. Taking the average of the panels' recommended cut scores allowed the overall MET writing test cut scores to reflect the judgments of both panels equally. It was feasible to do so because the panels' recommended cut scores were so close and because the panels showed good intra- and inter-panel consistency.

Table 3.6: Panel 1 Overall Cut Score Judgments

	Judgment Round 1			Judgment Round 2		
Judge ID	**A2/B1**	**B1/B2**	**B2/C1**	**A2/B1**	**B1/B2**	**B2/C1**
J1	18	27	37	18	27	37
J2	19	24	32	20	25	32
J3	19	25	33	19	25	33
J4	17	24	33	20	28	33
J5	17	26	38	17	26	38
J6	23	32	40	18	31	35
J7	20	28	35	18	28	35
J8	20	28	34	20	28	34
J9	20	30	37	20	30	33
Mean	19.22	27.11	35.44	18.89	27.56	34.44
Median	19	27	35	19	28	34
SD	1.856	2.713	2.698	1.167	2.068	2.007
SE_j	0.619	0.904	0.899	0.389	0.689	0.669
Min	17	24	32	17	25	32
Max	23	32	40	20	31	38

Table 3.7: Panel 2 Overall Cut Score Judgments

	Judgment Round 1			Judgment Round 2		
Judge ID	A2/B1	B1/B2	B2/C1	A2/B1	B1/B2	B2/C1
J1	20	27	33	20	28	33
J2	18	27	34	18	27	34
J3	25	29	34	20	28	34
J4	18	26	35	18	26	36
J5	17	27	31	17	27	33
J6	18	28	37	19	27	36
J7	21	29	34	21	29	34
J8	20	29	37	20	29	37
Mean	19.63	27.75	34.38	19.13	27.63	34.63
Median	19	27.5	34	19.5	27.5	34
SD	2.560	1.165	1.996	1.356	1.061	1.506
SEj	0.905	0.412	0.706	0.479	0.375	0.532
Min	17	26	31	17	26	33
Max	25	29	37	21	29	37

Table 3.8: Panel Cut Score Estimates

Panel	A2/B1	B1/B2	B2/C1
Panel 1	18.89	27.56	34.44
Panel 2	19.13	27.63	34.63
Overall	19.01	27.59	34.54

However, it is important to note that MET writing test scores consist of only integer numbers between 0 and 40. Therefore, the cut scores were rounded to 19, 28, and 35 for A2/B1, B1/B2, and B2/C1, respectively. While cut scores are generally rounded up to the nearest score point to minimize the chance of false positive classifications (Cizek & Bunch, 2007: 25), the raw cut score of 19.01 was rounded down to 19 for several reasons. First, it was felt that the ability level depicted by a score of 19.01 was so close to 19 that it was better represented by a score of 19 than 20. Second, it was decided that rounding the A2/B1 cut score down to 19 provided a better representation of the panelists' recommendations.

4. VALIDITY EVIDENCE

4.1 Procedural

This report has provided documentation of the panel composition, standard setting method, meeting procedures, and panelist results. Together, this documentation works to support the procedural validity of this study and the recommended cut scores. Here we provide additional procedural validity evidence by presenting the results from two questionnaires that were given to the panelists during the standard setting meeting. Both questionnaires collected data using a four-point Likert scale (1 – *Strongly Disagree* to 4 – *Strongly Agree*). The first questionnaire was given to the panelists upon completion of the familiarization and training tasks, but prior to the judgment task.

The results are summarized in Table 4.1. The table shows that the majority of the ratings were positive, and that the panelists understood the CEFR levels, the MET writing test, and the linking process. Only two panelists responded negatively, disagreeing with statements 6 and 7. One of these panelists felt that they did not have enough time to complete their individual tasks. This response probably referred to the familiarization task, since it was the most difficult and time consuming task. That said, this panelist's

Table 4.1: Results of Pre-Judgment Survey

No.	Question	Panel 1				Panel 2			
		1	2	3	4	1	2	3	4
1	The pre-reading helped me to understand the background to the CEFR.	-	-	5	4	-	-	4	4
2	The sample test helped me to understand the structure and level of the MET writing test.	-	-	2	7	-	-	1	7
3	The introductory presentation helped me to understand the linking process.	-	-	1	8	-	-	1	7
4	The discussion of pre-reading answered my questions.	-	-	1	7	-	-	2	6
5	The familiarization tasks helped me to understand the CEFR levels.	-	-	3	6	-	-	3	5
6	The training items helped me to understand the judgment process.	-	-	2	7	-	1	-	7
7	I had enough time to complete my individual tasks.	-	-	2	7	-	1	-	7
8	I had enough time to participate in the discussions.	-	-	1	8	-	-	2	6

responses to the other survey items suggests that his/her dissatisfaction with the time available to complete the individual tasks does not appear to have negatively affected his/her understanding of the CEFR or the judgment process. Another panelist indicated that the training items did not help him/her to understand the judgment process. However, it is important to note that this panelist indicated agreement with this same statement on the second questionnaire. This inconsistency seems to indicate an initial discomfort with the method, rather than a lack of understanding.

In addition to the statements summarized in the above table, the panelists were also asked if they were ready to proceed to the judgment task. All of the panelists indicated that they understood what was expected of them and that they were ready to proceed. After the judgment task, and at the conclusion of the standard setting meeting, the second questionnaire was administered to the panelists. The results are summarized in Table 4.2.

This table shows that, like the first questionnaire, the majority of the ratings were positive, and that the panelists understood all of the tasks performed during the meeting. Only one panelist responded negatively, disagreeing with statement 3. The panelist felt that they did not understand the instructions for each judgment round. However, this did not seem to impact their judgments, since the same panelist also indicated that they had confidence in their judgment decisions.

In addition to the above questions, panelists also responded to a question about the adequacy of the recommended cut scores (too low, about right, or too high). For Panel 1, seven panelists indicated that the recommended cut scores were about right. One of the panelists indicated that the cut scores were too high, while another indicated that only the B2/C1 cut score was too high, and that the others were about right.

Table 4.2: Results of Post-Judgment Survey

No.	Question	Panel 1				Panel 2			
		1	2	3	4	1	2	3	4
1	The familiarization tasks helped me to understand the CEFR levels	-	-	1	8	-	-	1	7
2	The training items helped me to understand the judgment process.	-	-	1	8	-	-	1	7
3	I understood the instructions for each judgment round.	-	-	2	7	-	1	-	7
4	I understood the group discussion of our judgments.	-	-	1	8	-	-	1	7
5	I had enough time to complete my individual tasks.	-	-	1	8	-	-	-	8
6	I had enough time to participate in the discussions.	-	-	1	8	-	-	-	8
7	I am confident in the decisions I have made.	-	-	3	6	-	-	2	6

For Panel 2, all eight of the panelists indicated that the recommended cut scores were about right. Overall, the panelists' responses and comments on the questionnaire were very positive, and indicate that they not only understood the entire standard setting procedure, but were also confident and happy with the resulting recommended cut scores.

4.2 Internal

The purpose of this section is to present internal validity evidence for the recommended MET writing test cut scores. The first piece of evidence is obtained by analyzing the cut scores recommended by the two independent panels. If the cut scores recommended by the panels are comparable, then the internal validity of the cut scores would be supported. Recall from Table 3.8 (in Section 3.3) that the recommended cut scores for the two panels were very similar, differing by less than half a score point for each. The statistical equivalence of these cut scores was examined using Welch's t-test, which tests the null hypothesis of equal means against the alternative of unequal means. Unlike the traditional two sample t-test, it does not assume the variances of the two populations are equal.

Table 4.3: Summary of Welch's T-Test

Results	A2/B1	B1/B2	B2/C1
t	-0.382	-0.089	-0.211
df	13.955	12.212	14.634
p	0.708	0.931	0.836

Table 4.3 summarizes the results (the test statistic [t], degrees of freedom [df] and p-value [*p*]) of Welch's t-test for each cut score. The table shows that the test fails to reject the null hypothesis of equal means for all three cut scores. This suggests that the two panels have provided statistically equivalent cut score recommendations.

Another piece of internal validity evidence is analysis of method consistency. This is examined using the standard error of judgment (SE_j). This measure offers "an estimate of the likelihood of replicating the recommended cut scores" (Tannenbaum & Cho, 2014). According to Cohen et al. (1999), the SE_j should be no more than half the standard error of measurement (SEM) of the test. The pilot MET writing test has an SEM of 2.097, which means that the SE_j should be less than 1.048 in order for an argument to be made for the validity of the panelists recommended cut scores. Tables 3.6 and 3.7 (in Section 3.3) present the SE_j values for the cut scores for the panels. For both panels, the SE_j of each cut score is much less than half of the SEM (ranging from 0.40 to 0.91). This suggests that the recommended cut scores are likely to be replicated if another standard setting study were performed.

The final piece of internal validity evidence presented for this study is decision consistency analysis. The consistency of the cut score decisions was examined using the methods and tables presented in Subkoviak (1988). It made use of the equation

$$|Z| = \frac{C - 0.5 - M}{S}$$

where C is the test's cut score, M is the mean of the observed test scores, and S is the standard deviation. Using the pilot MET writing test data as the observed data, this equation was applied to each of the three recommended cut scores. These values, along with the estimated reliability of the MET writing test, were used to obtain estimates of the agreement coefficient (p_0) and kappa (*k*) from the tables provided in Subkoviak (1988). It should be noted that linear interpolation was used to obtain p_0 and *k* values that better represented the reliability estimates. The results of this analysis are presented in Table 4.4 for each cut score.

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

Cut Score	p_0	*k*
A2/B1	0.926	0.639
B1/B2	0.851	0.692
B2/C1	0.947	0.619

When interpreting these statistics it is important to note that the agreement coefficient (p_0) is a measure of overall consistency, while kappa (*k*) is a measure of the test's contribution to that consistency (Subkoviak, 1988: 54). Because the maximum values in Subkoviak's tables are 0.98 for p_0 and 0.72 for *k*, it can be argued that the recommended MET writing test cut scores demonstrate acceptable decision consistency.

Overall, the three different pieces of internal validity evidence presented in this section provide support for the recommended MET writing test cut

scores. The Welch's t-test shows that two independent panels obtained cut scores that were statistically similar, while the small standard error of judgments indicate that the cut scores are likely to be replicated if another standard setting study were performed. The decision consistency analysis confirms that the MET writing test cut scores have acceptable decision consistency.

4.3 External

The purpose of this section is to present external validity evidence for the MET writing test's recommended cut scores. External validity evidence is often the hardest type of validity evidence to obtain, usually due to limitations of time or resources. The Council of Europe's manual to support CEFR standard-setting studies suggests collecting evidence from independent sources which support the outcome of the standard setting meetings (Council of Europe, 2009: Ch. 7). Examples would include analysis of test data for students who took both the MET writing test and another CEFR linked writing exam, or the use of a second standard setting method to verify the results. Unfortunately there was no information available about the ability level of the test takers for the pilot MET writing test and a second standard setting method is neither feasible nor recommended. Kaftandjieva (2010: 10) points out that, in addition to increasing the length and cognitive difficulty of the standard setting study, a second standard setting method would probably result in different cut scores and classification decisions.

Therefore, this report attempts to provide external validation through the exploration of the reasonableness of the cut scores. This was investigated through analysis of the MET writing test pilot data. The pilot population consisted of slightly different age and first language distributions than are currently typical for the MET but it was still representative of the target MET population, and can therefore provide useful insights.

The reasonableness of the MET writing test's cut scores was examined by comparing the CEFR distribution of the pilot data to the CEFR distributions of the 2013 test population for the MET listening and reading, sections and the MET speaking test. These distributions are presented in Table 4.5.

Table 4.5: CEFR Distribution of Pilot Candidates Based on the Recommended Cut Scores

MET Test	A2	B1	B2	C1
Listening	27.12	42.41	21.01	9.46
Reading & Grammar	25.22	46.23	21.90	6.61
Speaking	23.27	37.69	26.26	12.79
Writing	10.11	44.94	34.83	10.11

This table shows that the CEFR distribution of the MET writing test compares favorably to the other MET sections. The similarity in CEFR distribution helps to provide valuable external evidence to the validity of the MET writing test's recommended cut scores.

5. CONCLUSIONS

This technical report has presented the setting of the CEFR cut scores for the MET writing test. It has summarized the results of the standard setting study, along with the methodology used in obtaining these cut scores. It has also provided evidence of procedural, internal, and external validity in support of these cut scores. The final recommended CEFR cut scores for the MET writing test are presented in Table 5.1.

Table 5.1: Final MET Writing Test Cut Scores

A2/B1	B1/B2	B2/C1
19	28	35

Using these raw cut scores as a guide, all MET writing test scores will be reported on the MET standardized scale 0–80.

6. REFERENCES

Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental Measurements in the Human Sciences* (2nd ed.). Mahwah, N. J.: L. Erlbaum.

Brunfaut, T. & Harding, L. (2014). *Linking the GEPT Listening Test to the Common European Framework of Reference.* Taiwan: Language Training and Testing Centre.

CaMLA (2014). *Linking the Common European Framework of Reference and the CaMLA English Placement Test*, CaMLA Technical Report, CaMLA. Retrieved from http://www.cambridgemichigan.org/sites/default/files/resources/Reports/EPT-Technical-Report-20140625.pdf.

Cizek, G. J. & Bunch, M. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* London: Sage Publications.

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.

Generalitat de Catalunya (2006). *Proficiency scales: The Common European Framework of Reference for Languages in the Escoles Oficials d'Idiomes in Catalunya*. Madrid: Cambridge University Press.

Kaftandjieva, F. & Takala, S. (2002). Council of Europe scales of language proficiency: A validation study. In J. C. Alderson (Ed.), *Common European Framework of Reference for Languages: Learning, teaching, assessment. Case studies.* (pp. 106–129). Strasbourg: Council of Europe.

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.

Kane, M. (2001). So much remains the same: conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53–88). Mahwah, N.J.: Lawrence Erlbaum Associates.

Kane, M., Crooks, T., & Cohen, A. S. (1999). Validating measures of performance. *Educational measurement: Issues and Practice, 18*(2), 5–17.

Lewis, D. M., Mitzel, H. C., Mercado, R. L., & Schulz, E. M. (2012). The Bookmark Standard Setting Procedure. In G. J. Cizek (Ed.), *Setting performance standards: foundations, methods, and innovations* (2nd ed., pp. 225–253). New York, N.Y.: Routledge.

Linacre, J. M. (2002). What do Infit and Outfit mean-Square and Standardized mean? *Rasch Measurement Transactions*, *16*(2), 878.

Morrow, K. (2004). Background to the CEF. In K. Morrow (Ed.), *Insights from the Common European Framework*. (pp.3–11). Oxford: OUP.

North, B. (2000). *The development of a common framework scale of language proficiency*. New York: Peter Lang.

North, B. & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing*, *15*(2), 217–262.

Papageorgiou, S. (2007). *Relating the Trinity College London GESE and ISE examinations to the Common European Framework of Reference: Final Project Report, February 2007*, Trinity College London. Retrieved from http://www.trinitycollege.co.uk/resource/?id=2261.

Papageorgiou, S. (2010). *Setting cut scores on the Common European Framework of Reference for the Michigan English Test*, CaMLA Technical Report, CaMLA. Retrieved from http://www.cambridgemichigan.org/sites/default/files/resources/MET_StandardSetting.pdf.

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. & Katz, I. R. (2008). *Setting Standards on the Core and Advanced iSkillsTM Assessments*, ETS RM-08-04, Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RM-08-04.pdf.

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.

Zieky, M. J., Perie, M., & Livingston, S. A. (2008). *Cutscores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.