met michigan english test

Setting Cut Scores on the Common European Framework of Reference for the Michigan English Test

Technical Report

Setting Cut Scores on the Common European Framework of Reference for the Michigan English Test

Technical Report

Spiros Papageorgiou

Testing and Certification Division
University of Michigan
English Language Institute

500 East Washington Street
Ann Arbor, MI 48104-2028

www.lsa.umich.edu/eli/testing/met
michengtest@umich.edu

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

The 56 Reading descriptor statements for the CEFR familiarization activities were used with the kind permission of Dr. Felianka Kaftandjieva. Many thanks go to Dr. Jayanti Banerjee for commenting on an earlier version of this report.

The Bi-national Centers in Colombia (known as "Centro Colombo Americano," or simply "Colombo") have contributed immensely by authorizing the participation of their employees as judges. This allowed for representation of all nine locations where the MET pilot test was taken in 2008 and where the operational test was administered throughout 2009. Special thanks go to the Bi-national Center in Cali for organizing the meeting and offering facilities to conduct it.

1. INTRODUCTION

This report presents the results of a standard-setting project to set cut scores on the proficiency levels of the Common European Framework of Reference (CEFR, Council of Europe, 2001) for the Michigan English Test (MET). The methodology and the results of this project are discussed.

1.1 STANDARD SETTING

Standard setting is defined as the decision-making process of classifying examinees into a number of levels or categories such as "advanced," "proficient," "basic," and "below basic" (Kane, 2001: 53). The "boundary between adjacent performance categories" (Kane et al., 1999: 344) is defined by a cut score. In other words, a cut score is "a point on a test's score scale used to determine whether a particular score is sufficient for some purpose" (Zieky et al., 2008: 1). To determine for example whether examinees have passed or failed an exam, the cut score functions as the boundary between the pass and fail category.

During a standard-setting meeting, a panel of expert judges (commonly referred to as panelists) is required to make judgments on which examination providers will base their final cut score decisions. Under the guidance of one or more meeting facilitators, statistical information about test items and the distribution of scores are provided to help panelists with their judgment task. More than one round of judgments is usually organized to allow judges to discuss their decisions, take into account the relevant statistical information, and revise their judgments.

As presented in the revised version of the Council of Europe's Manual (2009: Ch. 7) the standard-setting meeting must be evaluated in terms of three main categories:

- **Procedural validity**, examining whether the procedures followed were practical, implemented properly, that feedback given to the judges was effective and that documentation has been sufficiently compiled.
- **Internal validity**, addressing issues of accuracy and consistency of the standard setting results.
- **External validation**, by collecting evidence from independent sources which support the outcome of the standard setting meeting.

1.2 THE COMMON EUROPEAN FRAMEWORK

The publication of the CEFR has been recognized as the "most significant recent event on the language education scene in Europe" (Alderson, 2005: 275). The CEFR scales and their constituent descriptors, developed during a large research project (North, 2000a; North & Schneider, 1998), describe what learners can do with language at six main levels (A1, the lowest, to C2, the highest). When setting cut scores, these descriptors can function as "Performance Level Descriptions" (PLD) and the level names (A1, A2, etc.) are the summarizing labels of these descriptions, called "Performance Level Labels" (PLL; see Cizek & Bunch, 2007: 44–47).

1.3 THE MANUAL FOR RELATING EXAMINATIONS TO THE CEFR

To assist test developers in relating their examinations to the CEFR, the Council of Europe has published a preliminary—and recently, a revised version—of its manual (Council of Europe, 2003, 2009; Figueras et al., 2005) that includes suggested standard-setting procedures. A reference supplement (Takala, 2004) has also been issued, with one section focusing on standard setting (Kaftandjieva, 2004).

1.4 THE MICHIGAN ENGLISH TEST

The Michigan English Test (MET) is a standardized, multi-level examination of general English language proficiency provided by the University of Michigan English Language Institute (ELI-UM). It measures listening, reading, grammar, and vocabulary skills in personal, public, occupational, and educational contexts. Listening recordings and reading passages reflect everyday, authentic interaction in a North American English linguistic environment. 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 MET can be used for educational purposes, such as when finishing an English language course, or for employment purposes, such as applying for a job or pursuing promotion that requires an English language qualification.

The MET consists of 135 multiple-choice questions, with four answer options per item. Section I: Listening contains 60 items of three types: short dialogues with one question, longer dialogue sets preceding three to four questions, and monologue sets followed by four to five questions. Section II: Grammar and Reading contains 25 grammar questions and 50 reading comprehension

items. Vocabulary is tested in listening and reading items. Test takers receive a scaled score in the 0–80 range for each section and a final score which is the total of the two section scores.

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

1.5 PURPOSE FOR SETTING CUT SCORES ON THE CEFR LEVELS

The setting of the MET cut scores on the CEFR levels was decided by ELI-UM so that exam results are more meaningful to test users, given that the CEFR levels are widely used worldwide to interpret test scores. In order to identify cut scores, a standard setting meeting took place in Cali, Colombia, with participation of judges from the Bi-national Centers (BNCs) and a facilitator from ELI-UM.

2. METHODOLOGY

The methodology of the study is chronologically presented in this section (i.e., before, during and after the meeting).

2.1 SELECTION OF JUDGES

The panel of judges consisted of 13 participants from all nine BNCs in Colombia, which administer the MET. Two criteria were used to select the standard setting panel judges:

i. Knowledge of the test-taking population

ii. Geographical coverage of all test locations

The latter criterion was important because information about item difficulty during the standard setting meeting was based on the pilot administration of the MET in all nine centers. The final panel comprised 13 participants from all nine BNCs in Colombia where the MET is administered. Their positions were varied and included English teachers and teacher trainers, academic directors, and academic advisors.

However, the level of the judges' familiarity with the CEFR could not be established prior to the meeting; thus, familiarization activities were added to the meeting program, based on Chapter 3 of the Manual (Council of Europe, 2009). Due to security reasons, all judges signed confidentiality agreements that they would not release information about items.

2.2 STANDARD-SETTING METHOD

The standard-setting method was based on that proposed by Angoff (Angoff, 1971), probably the most frequently used and well-researched method in the last four decades (cf., Zieky et al., 2008: 62–63). The Angoff method was preferred because of the clarity of the judgment task and the flexibility in terms of organizing the judgment session.

The judges were asked to think of 100 B1 borderline examinees—that is, 100 examinees that had just passed the border between A2 and B1. For each item, the judges were asked to state how many of these 100 examinees would answer each item correctly (cf., Cizek & Bunch, 2007: 85). The same question was repeated for 100 B2 and 100 C1 borderline learners for each item of the two MET Sections (i.e., examinees that had just passed the border between B1 and B2 and B2 and C1 levels respectively).

It should be noted here that the judgment task of the Angoff method is commonly phrased as the probability of an imaginary, borderline examinee answering an item correctly (Angoff, 1971: 515). However, asking judges to estimate the number of successful examinees out of 100 is seen in the literature as a more accessible task, due to the difficulty that judges may have in understanding the notion of probability.

2.3 MATERIAL

In preparation for the standard-setting meeting, material to familiarize the judges with the CEFR levels was prepared. Fifty-six reading, 71 listening, 17 grammar and 25 vocabulary sentence-level statements from the CEFR descriptors (see sample in Appendix 1) were presented to the judges asking them to choose the CEFR level they belong to (A1-C2). No indication of the level was presented to the judges. For faster analysis of results, the judges were asked to use numbers instead of levels in the following way: A1-1; A2-2; B1-3; B2-4; C1-5; and C2-6.

The "atomization" of the descriptors into short statements, based on Kaftandjieva and Takala (2002), aimed to familiarize the judges with all constituent statements of the descriptors, which usually contain a number of sentence-level statements. It would be

reasonable to claim that the task became more difficult for the judges, because a number of statements were quite short, without a detailed description of the context of language use (e.g., L01 in Appendix 1).

In order to help judges obtain a better understanding of the difficulty of test items and how this relates to the judgment task, the training material asked judges to rank a number of listening and reading MET pilot items from easiest to most difficult (Appendix 2).

Finally, using the Angoff method (Section 2.2) the judges were asked to estimate for each item the number of 100 test takers at the border between two CEFR levels that would answer the item correctly. Each judge's estimates were then added and divided by 100, which led to three proposed cut scores (A2/B1, B1/B2, and B2/C1 borders) by each judge in the form of total number of items answered correctly (see sample rating form in Appendix 3).

2.4 TASKS DURING THE MEETING

The first day of the meeting was dedicated to CEFR familiarization and item difficulty training tasks. The judges worked individually in guessing the level of the CEFR descriptors and the results were analyzed by the moderator overnight so that a discussion could take place during the following two days. The same applied to the item difficulty task.

The following two days were structured similarly, with reading, grammar, and vocabulary preceding listening. Firstly, the moderator used a projector to present the familiarization activity results on an Excel spreadsheet. When a descriptor statement had a median of judgments that did not agree with the correct CEFR level and/or the range of judged levels was 3 and above, the moderator invited participants to explain the reasons for choosing a particular level. Moving to the next descriptor was done only after all judges felt they understood the correct CEFR level of a descriptor statement. To avoid embarrassing any judges, each participant was given an envelope with an enclosed alphanumeric ID (J1, J2, etc.) so that someone's identity could not be revealed by looking at the spreadsheet.

The CEFR familiarization task was followed by the discussion of the item difficulty task. The discussion led by the moderator aimed to demonstrate that predicting item difficulty is not always successful, as well as to link the cut score judgment task to the notion of item difficulty (i.e., the more difficult an item the higher the CEFR level). The moderator also explained the notion of facility value and point-biserial correlation, placing emphasis on the relationship between the facility value and the number of examinees that the judges predict will answer the item correctly (i.e., the higher the facility value figure, the easier the item and thus the higher the predicted number of examinees answering the item correctly should be).

The cut score judgment task was organized in two rounds, which is a fairly common practice in standard setting meetings (Hambleton, 2001). Initially the judges received a booklet and a rating form (see sample in Appendix 3). They first took the test and answered each item and then they inserted their judgments in the rating form. Each judge used an electronic calculator to report the last line of the rating form (TOTAL/100) to the moderator, which, as mentioned earlier, resulted in three cut scores (A2/B1, B1/B2 and B2/C1) in terms of total items answered correctly. These appeared on the projected Excel spreadsheet for discussion.

After presenting, discussing each judge's recommended cut scores and summarizing these in terms of mean, median, maximum and minimum cut scores, the moderator provided further feedback in the form of a statistical analysis from an MET pilot administration in July 2008. The judges received a two-page document on each day for one the two MET sections (Section I: Listening and II: Grammar and Reading) which contained the following information:

- Number of test takers (N)
- Number of items (k)
- Mean score
- Standard deviation
- Maximum score
- Minimum score
- Cronbach's (alpha)
- Facility value and point-biserial correlation for each item
- Histogram showing the distribution of scores

Statistics were first explained and the judges were asked to consider them, if they wanted to, before they repeated the same cut score task, using a rating form that was identical to the one for Round 1. Judges were reminded that they were allowed to change their estimates or simply keep the same estimates from the previous round. The Round 2 judgments were presented on the spreadsheet for any final comments. It should be noted that J12 could not attend the meeting on the final day. Thus, cut score estimates for Section I were provided by 12 judges (rather than 13).

2.5 POST-MEETING ANALYSIS OF DATA

After the standard setting meetings, the data was inspected for accuracy (Davidson, 1996) and the cut score judgments were examined for inter-judge and intra-judge consistency (see Section 4).

3. RESULTS OF THE CEFR FAMILIARIZATION ACTIVITIES

Before the analyses of the cut scores, the familiarization activities will be discussed first, to establish the judges' familiarity with the CEFR levels. If judgments are made by participants who do not have a good understanding of the CEFR levels, this will cast doubt on the validity of the recommended cut score because, when the judges recommend cut scores, the underlying premise is that they fully and completely understand the CEFR levels and can rank the CEFR descriptors in the correct order and also assign random descriptors to the correct CEFR level. If they cannot assign descriptors consistently to the correct CEFR level then they are likely to provide inconsistent judgments when setting cut scores.

The analysis examined the number of descriptors that were placed at the correct level. Moreover, correlations of the judges' level placements and the correct levels were run. High correlations indicate that the judges understand how the descriptors progress from lower to higher levels. However, correlations do not show how many descriptors were placed at the correct level, thus they should be consulted along with the number of correct level placements. Finally, each judge's mean level was calculated to establish judges' tendency to put descriptors at lower or higher levels. The four tables below (Table 3.1 to Table 3.4) present information about the judges' performance during the familiarization task. Similar analysis can be found in other relevant studies, such as Kaftandjieva and Takala (2002), the Ministry of Education in Catalonia, Spain (Generalitat

Table 3.1 Listening Familiarization Task Results (71 descriptors, mean level 3.56)

	J1	J2	J3	J4	J5	J6	J7	J8	J9	J10	J12	J13	J14
Correct	41	41	22	31	32	24	26	43	28	19	36	29	27
Spearman	0.85	0.89	0.80	0.81	0.71	0.77	0.79	0.88	0.80	0.70	0.91	0.84	0.79
Mean	3.73	3.79	3.69	3.72	3.86	4.2	4.37	3.55	3.92	4.56	3.66	4.10	3.65

Table 3.2 Reading Familiarization Task Results (56 descriptors, mean level 3.25)

	J1	J2	J3	J4	J5	J6	J7	J8	J9	J10	J12	J13	J14
Correct	37	29	32	30	19	24	12	23	23	16	28	25	27
Spearman	0.92	0.92	0.85	0.86	0.69	0.86	0.84	0.84	0.82	0.62	0.90	0.86	0.77
Mean	3.20	3.50	3.59	3.48	3.77	3.66	4.29	3.38	3.46	4.02	3.20	3.88	3.36

Table 3.3 Vocabulary Familiarization Task Results (25 descriptors, mean level 3.40)

	J1	J2	J3	J4	J5	J6	J7	J8	J9	J10	J12	J13	J14
Correct	12	14	12	20	13	12	10	16	15	8	21	17	11
Spearman	0.89	0.93	0.91	0.96	0.70	0.76	0.73	0.92	0.90	0.84	0.97	0.90	0.86
Mean	3.00	3.20	3.20	3.28	3.28	3.84	3.48	3.44	3.20	4.16	3.32	3.28	2.92

Table 3.4 Grammar Familiarization Task Results (17 descriptors, mean level 3.41)

	J1	J2	J3	J4	J5	J6	J7	J8	J9	J10	J12	J13	J14
Correct	7	10	9	12	10	4	5	13	9	5	6	5	5
Spearman	0.90	0.94	0.97	0.87	0.91	0.95	0.89	0.95	0.84	0.78	0.93	0.85	0.89
Mean	3.94	3.82	3.88	3.53	3.76	4.24	3.88	3.65	3.47	4.24	3.94	3.71	3.94

de Catalunya, 2006), and the Trinity College London CEFR Project Report (http://www.trinitycollege.co.uk/resource/?id=2261). The first row shows the number of descriptors placed at the correct level by each judge (J1 to J14; no J11 ID used as it was reserved initially for a judge who could not attend). The second row presents the Spearman correlation between each judge's descriptor placements and the correct CEFR levels. The correlation was calculated by comparing a judge's level placements with the correct ones. The last row presents the mean level of all level choices by each judge which can then be compared to the CEFR mean level indicated in the parentheses of the caption of each table. If a judge's mean is higher than the CEFR mean, then this indicates a tendency to assign descriptors to a higher level than the correct one. When a judge's mean is lower than the CEFR mean, the tendency to assign descriptors to a lower level than the correct one is probably the case. Such a tendency to assign descriptors to a lower or higher CEFR level might result in inconsistent judgments when setting cut scores.

The generally high correlations suggest that the judges had in general a good understanding of how language proficiency progresses from lower to higher CEFR levels. All correlations were statistically significant ($p \leq 0.01$). However, correlations, as Kaftandjieva (2004: 23) points out, may mask low exact agreement of ratings. The low number of correct placements in the first row points to problems the judges had with the exact level. Eyeballing the raw data confirms the tendency highlighted by each judge's mean level in the third row; if this level is compared to the CEFR mean level in the caption, the judges tend to be lenient, with the exception of the Vocabulary descriptors. Lenient judges in this context tend to place descriptors at a higher level, for example a B1 descriptor at B2, a B2 descriptor at C1 and so on. The implication for setting cut scores is important, as judges might transfer leniency or severity from the familiarization task to their cut score judgments. The judges of this study demonstrate some leniency during the familiarization task, which might result into lower cut scores. Action taken in the meeting to avoid this will be discussed towards the end of this section.

So far the analysis examined individual judges' understanding of the CEFR levels. However, cut scores are usually based on the judgments of the panel, not just an individual judge. Therefore, further analysis, presented in Table 3.5, was performed to establish the consistency of the panel of judges. Cronbach's (alpha) is an internal consistency index usually used in item-based tests, but following the aforementioned studies analyzing similar familiarization tasks, it is reported here to indicate "the consistency of the reliability of ratings in terms of rater consistency" (Generalitat de Catalunya, 2006: 62). The intraclass correlation coefficient, ICC (Generalitat de Catalunya, 2006: 62), is calculated in order to demonstrate how the average rater agreed with all others. Nichols' (2006) guidelines on how to calculate ICC with SPSS were followed. A more detailed discussion of ICC can be found in McGraw and Wong (1996). The ICC two-way mixed model was used and average measures for exact agreement are reported. Kendall's W was also used to investigate rater agreement in similar contexts (Generalitat de Catalunya, 2006: 112; Kaftandjieva & Takala, 2002). As the SPSS Help Tool explains, Kendall's W can be interpreted as a coefficient of agreement among raters. Each case (row) is a rater and each variable (column) is an item being rated. The coefficient W ranges from 0 to 1, with 1 indicating complete inter-rater agreement, and 0 indicating complete disagreement among raters. All indices show high consistency and agreement among judges.

To conclude, the analysis suggests that the panel is consistent and has an overall good understanding of how language ability progresses from lower to higher levels in the CEFR scales. Nevertheless, the judges had issues with placing the descriptors at the exact level. Even though occasionally placing a B2 descriptor at C1 or B1 is not unreasonable, judges who possibly misunderstand some of the differences between adjacent CEFR levels will probably recommend cut scores whose validity should be

Table 3.5 Agreement and Consistency of the Group (all familiarization tasks)

	Reading	Listening	Grammar	Vocabulary
ICC	0.98	0.98	0.98	0.98
W	0.81	0.81	0.84	0.79
Alpha	0.98	0.98	0.98	0.98

questioned. In particular, the panel of this study might recommend cut scores that are lenient. To avoid this, the following steps were followed:

- Data were analyzed overnight to identify leniency or severity and to establish consistency.
- On each of the two following days of the meeting, the moderator presented the relevant task results to the judges and, using descriptive statistics, disagreement about the level and tendencies with regard to leniency/severity were highlighted.
- Each descriptor statement was discussed within the group and, with the moderator's guidance, the group agreed on the content of each statement that clearly indicated the level.

The judges pointed out that the discussion was very useful, as it helped them clarify the differences between adjacent levels and look at the descriptors more carefully. Even though there is no empirical evidence to confirm the positive impact of the discussion (e.g., by repeating the task), one would expect that the steps taken to explain the familiarization results and give feedback to the judges were in the right direction. Finally, to further support the judges during their subsequent cut score tasks, the descriptor statements were ordered in a different handout not according to their numeric ID, but by the level they belong to. This handout was used by the judges as the set of Performance Level Descriptions according to which they should make their cut score decisions.

4. CUT SCORE RESULTS AND VALIDITY EVIDENCE

4.1 CUT SCORE VALIDATION

Standard setting validation includes three main areas as illustrated in the figure below.[1] Arguments supporting procedural validity have already been presented, as the methodology of this standard setting study has been documented in detail in the previous section and was based on recommendations in the relevant literature. This section will be concerned with the resulting cut score and evidence of internal and external validation. It will also discuss the judges' feedback as part of procedural validation.

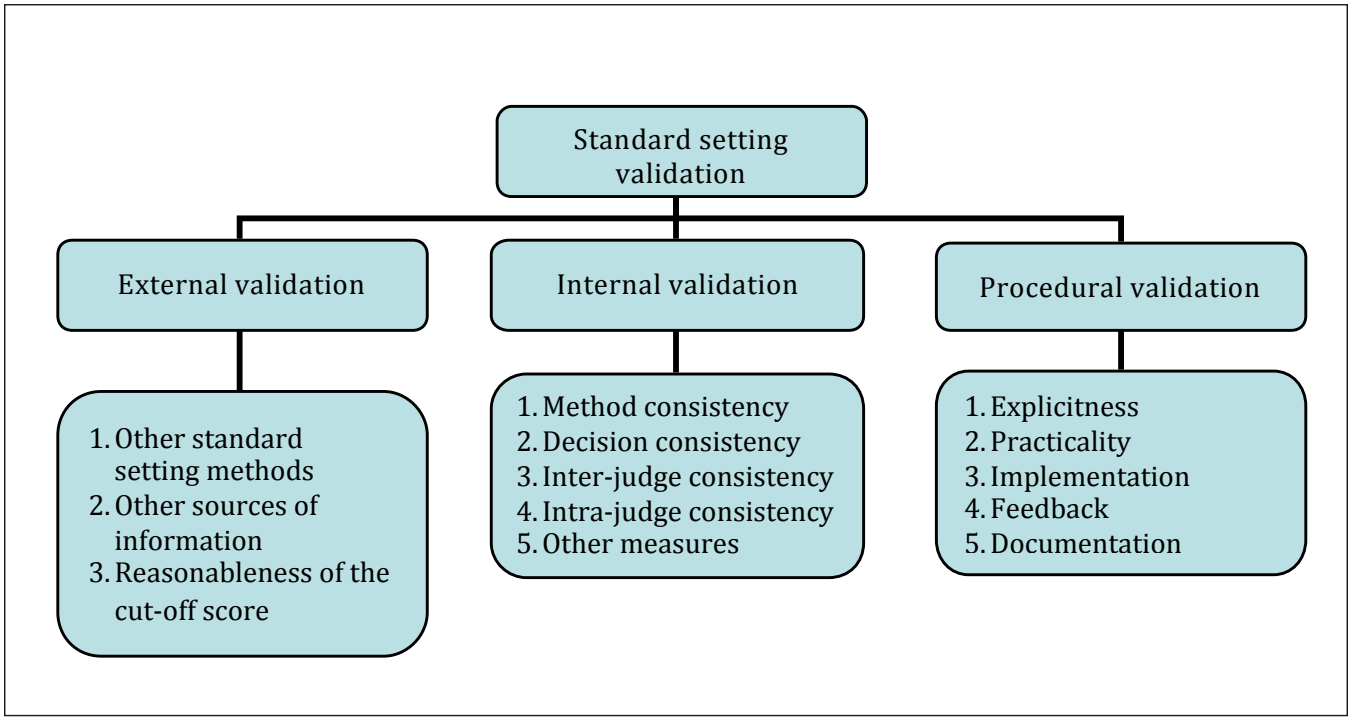

Figure 4.1: Standard-Setting Validation Areas

1 The figure is based on the pre-EALTA conference standard setting workshop in Barcelona, Spain (by Figueras, Kaftandjieva and Takala) and the 2009 version of the Manual (Council of Europe, 2009).

4.2 INITIAL CUT SCORE ESTIMATES

Table 4.1 presents the cut score judgments for Section I and Table 4.2 the same information for Section II. Both tables show the three cut scores each judge recommended in both rounds in terms of number of items correct (60 for Section I and 70 for Section II[2]). Statistics of these judgments are summarized in the lower part of the tables. The Round 2 cut scores are the final ones, because they were made after judges compared their cut scores to those recommended by other judges as well as the item analysis data. The recommended cut scores were as follows, as can be seen by the mean values of the Round 2 judgments:

- Section I: B1-18; B2-37; C1-49
- Section II: B1-22; B2-44; C1-58

It should be noted that numbers were rounded up, in order to minimize any false positive classifications (Cizek & Bunch, 2007: 25), given the high-stakes nature of the test. For example the C1 cut score for Section II was 57.45 and rounding to the nearest whole number would result into a cut score of 57. However, examinees who answer 57 items correctly do not demonstrate the ability depicted by a cut score of 57.45. Since the number of correctly answered items can either be 57 or 58, 58 was the cut score chosen. A number of cut score validation analyses were then run as illustrated below: method and decision consistency and intra-judge and inter-judge consistency.

Table 4.1 Cut Score Judgments for MET Section I (Listening)

Judge ID	Round 1 B1	Round 1 B2	Round 1 C1	Round 2 B1	Round 2 B2	Round 2 C1
J1	19.25	33.70	47.65	21.10	35.60	48.50
J2	13.55	31.40	46.80	21.50	39.10	55.70
J3	12.50	27.90	44.90	14.20	30.50	47.80
J4	26.39	40.58	51.18	31.00	41.40	48.80
J5	23.80	46.20	54.00	13.00	36.50	49.50
J6	11.70	35.70	52.70	12.30	33.30	51.30
J7	10.51	41.86	55.39	15.60	41.80	56.30
J8	11.50	33.00	51.81	12.30	34.80	49.00
J9	10.15	34.90	45.98	10.20	34.90	46.00
J10	19.05	25.75	29.10	24.10	31.20	35.70
J13	12.05	33.10	46.90	9.85	29.40	41.90
J14	17.00	43.15	54.75	24.90	43.70	54.90
Mean	15.62	35.60	48.43	17.50	36.01	48.80
SD	5.44	6.21	7.10	6.82	4.65	5.84
Min	10.15	25.75	29.10	9.85	29.40	35.70
Max	26.39	46.20	55.39	31.00	43.70	56.30

2 The pilot form used in the study contained five fewer items in Section II. This was taken into account when finalizing the Section II cut scores at ELI-UM (see Section 5)

Table 4.2 Cut Score Judgments for MET Section II (Reading and Grammar)

Judge ID	Round 1 B1	Round 1 B2	Round 1 C1	Round 2 B1	Round 2 B2	Round 2 C1
J1	13.90	33.00	57.45	19.53	37.30	60.20
J2	27.60	51.00	65.95	25.90	48.40	65.30
J3	33.95	55.20	63.70	21.80	39.60	58.60
J4	25.85	42.30	55.90	20.13	31.50	41.03
J5	17.75	42.30	62.30	27.20	48.30	63.10
J6	23.35	47.80	62.65	15.70	50.40	65.90
J7	16.27	46.65	63.32	17.98	45.30	63.32
J8	2.90	16.50	38.80	15.40	34.84	58.00
J9	28.05	46.70	60.05	32.05	46.75	57.70
J10	48.50	60.70	66.70	32.90	40.50	41.45
J12	18.10	34.30	50.06	15.98	29.98	44.28
J13	38.95	57.30	65.55	35.75	51.60	59.70
J14	17.70	48.45	66.95	30.60	57.10	68.30
Mean	24.07	44.79	59.95	23.92	43.19	57.45
SD	11.82	11.72	8.00	7.20	8.28	9.27
Min	2.90	16.50	38.80	15.40	29.98	41.03
Max	48.50	60.70	66.95	35.75	57.10	68.30

4.3 METHOD CONSISTENCY ANALYSIS AND FINALIZATION OF CUT SCORES

The first analysis examined **method consistency** by estimating the standard error of judgment (SEj). This is calculated by dividing the standard deviation of judgments with the square root of the number of judges (Norcini et al., 1987). According to Cohen et al. (1999), SEj should be equal to or smaller than half of the standard error of measurement (SEM) of the test. MET Form A, which was judged here, has a SEM of 3.42 for the 60-item Listening Section and a SEM of 3.49 for the 70-item Reading and Grammar Section. Thus, in order to argue for the validity of the cut score, the SEj should be equal to or smaller than 1.71 and 1.74 for each section respectively. However, as can be seen in the second column of Table 4.3, this is only the case for two cut scores (rows 2 and 3). In order to minimize the SEj, extreme ratings (too low and too high cut scores) were excluded for the four cut scores until the SEj was below

Table 4.3 Comparison of SEj Before and After Excluding Extreme Ratings

Cut score	SEj incl. extreme ratings	SEj excl. extreme ratings (if necessary)
Section I B1	1.97	1.57
Section I B2	1.34	1.34
Section I C1	1.69	1.69
Section II B1	2.00	1.71
Section II B2	2.30	1.62
Section II C1	2.57	1.71

Table 4.4 Recommended Cut Score for MET Section I (Listening)

Judge ID	B1	B2	C1
J1	21.10	35.60	48.50
J2	21.50	39.10	55.70
J3	14.20	30.50	47.80
J4	(excl.)	41.40	48.80
J5	13.00	36.50	49.50
J6	12.30	33.30	51.30
J7	15.60	41.80	56.30
J8	12.30	34.80	49.00
J9	10.20	34.90	46.00
J10	24.10	31.20	35.70
J13	(excl.)	29.40	41.90
J14	24.90	43.70	54.90
Mean	16.90	36.01	48.80
SD	5.44	4.65	5.84
Min	10.20	29.40	35.70
Max	24.90	43.70	56.30

1.71 and 1.74 (see also calculation of the trimmed mean in Zieky et al., 2008: 38–39). The mean values in Table 4.4 for Section I and Table 4.5 for Section II suggest the following cut scores after the exclusion of the extreme ratings (indicated in the tables):

- Section I: B1-17; B2-37; C1-49
- Section II: B1-25; B2-45; C1-60

Lowering the SEj is expected to result in more valid cut scores, as extreme judgments are excluded. It should also be pointed out that all Section II cut scores (Table 4.5) were raised compared to those mentioned earlier, which is likely to be a good decision for two reasons: first, the group tended to be relatively lenient according to the results of the familiarization activities. Second, the high-stakes nature of the exam dictates special caution when it comes to false positive classifications, which are likely to be more important than false negative classifications.

4.4 DECISION CONSISTENCY ANALYSIS

After excluding the extreme ratings, **decision consistency** (Cizek & Bunch, 2007: 307) was examined using the following equation:

$$|Z|=(C_X - M - 0.5)/S_X$$

where C_X is the cut score for the test, M is the observed test mean and S_X is the standard deviation (SD) of observed test scores. Absolute values of Z are then used to obtain the estimates of agreement coefficient (p_0) and kappa (k) from two tables in Subkoviak (1988), reproduced in Cizek and Bunch (2007: 310–311). Table 4.6 presents the results for Section I (Listening; mean 27.83; SD 11.61) and Section II (Reading; mean 42.18; SD 12.65). In Subkoviak's table, the maximum value for p_0 is 0.98 and 0.71 for k. It could therefore be argued that the MET recommended cut scores for CEFR levels B1, B2 and C1 demonstrate satisfactory decision consistency.

Table 4.5 Recommended Cut Score for MET Section II (Reading and Grammar)

Judge ID	B1	B2	C1
J1	19.53	37.30	60.20
J2	25.90	48.40	65.30
J3	21.80	39.60	58.60
J4	20.13	(excl.)	(excl.)
J5	27.20	48.30	63.10
J6	(excl.)	50.40	65.90
J7	17.98	45.30	63.32
J8	(excl.)	34.84	58.00
J9	32.05	46.75	57.70
J10	32.90	40.50	(excl.)
J12	15.98	(excl.)	44.28
J13	(excl.)	51.60	59.70
J14	30.60	(excl.)	(excl.)
Mean	24.41	44.30	59.61
SD	6.15	5.82	6.15
Min	15.98	34.84	44.28
Max	32.90	51.60	65.90

Table 4.6 Agreement Coefficient (p_0) and Kappa (k) for the MET Cut Scores

Cut score	p_0	k
Section I B1	0.90	0.68
Section I B2	0.88	0.70
Section I C1	0.97	0.61
Section II B1	0.95	0.64
Section II B2	0.86	0.71
Section II C1	0.94	0.65

Table 4.7 Correlations Between Mean of Judgments and Empirical Difficulty

	Section I – Round 1	Section I – Round 2	Section II – Round 1	Section II – Round 2
Spearman rho	0.42	0.83	0.73	0.92

4.5 INTRA-JUDGE AND INTRA-JUDGE CONSISTENCY

Additional analysis examined **intra-judge** and **inter-judge consistency**. The first was examined by correlating the mean ratings of the judges (without the excluded extreme ones) with the empirical difficulty. All correlations (Table 4.7) were significant ($p \leq .01$). Given that estimating difficulty is a hard task for expert judges (cf. Alderson, 1993), correlations above 0.30 are considered satisfactory. This was achieved with both MET Sections. As in the case of the familiarization activities, statistics of agreement and consistency of the group were analyzed (Table 4.8). These were once again found to be high, offering additional cut score validity evidence.

Table 4.8 Agreement and Consistency of the Group (cut score tasks)

	Section I	Section II
ICC	0.94	0.94
W	0.80	0.76
Alpha	0.94	0.94

4.6 EXTERNAL VALIDATION

The analyses presented in Sections 4.3 and 4.4 (method and decisions consistency and intra-judge and inter-judge consistency) offer some evidence in terms of internal validation (Figure 4.1). To obtain some external validation evidence, the Manual suggests the collection of evidence from independent sources which support the outcome of the standard setting meeting (Council of Europe, 2009: Ch.7). For example, the same test takers could take another test already calibrated to the CEFR. Alternatively, a second standard setting method could be used. Unfortunately, it was it was not possible to collect such information. Asking examinees to take a second examination on the day they took Pilot Form A was logistically impossible, as it would increase seat time and fatigue. It was also very difficult to choose an examination aiming at the same CEFR levels, intended for the same purposes and tapping a similar construct. Moreover, a second method would increase the judges' cognitive load, and fatigue could potentially affect their judgments with the already selected method. It should be pointed out that a whole day was dedicated to familiarization. This was essential, given that the judges' familiarity with CEFR was unclear; as a result, there was no time for using a second standard setting method. Thus, as it will be shown below, external validation was attempted by exploring the reasonableness of the cut scores and by comparing how examinees were classified into levels based on the cut scores of this study and a test center's independent classification.

The **reasonableness of the cut scores** was investigated by showing the judges how these cut scores would group the 660 test takers that took MET Form A into levels. Results are illustrated in Table 4.9 for both MET sections, Figure 4.2 for Section I and Figure 4.3 for Section II. In the group discussion that followed, all judges felt that the recommended cut score yielded reasonable classifications of test takers. It should be reminded that, apart from the familiarization activities and the feedback in terms of empirical evidence during the cut score task, the judges also took the test so that they have a better understanding of the test takers' experience.

Table 4.9 Classification of Form A Test Takers (N = 660) into CEFR Levels Based on the Recommended Cut Scores

	Section I	Section II
A2	105 (15.91%)	55 (8.33%)
B1	408 (61.81%)	323 (48.94%)
B2	95 (14.39%)	214 (32.43%)
C1	52 (7.88%)	68 (10.30%)

Comparison of level classifications was also explored as additional piece of external validity evidence based on North (2000b), who used teacher judgments to verify the cutoffs for items stored in a bank (see also Council of Europe, 2009: 111). The academic director of one of the centers, who also acted as a judge, was asked to provide estimates of CEFR level for the 302 students of her center who took the MET Form A (used in this standard setting study), based on the classes they attended at the center. These students were also classified into CEFR levels based on the number of Form A items they answered. Thus it was possible to compare how students were classified into levels based on two different sources (test center and test items).

Table 4.10 shows a moderate Pearson correlation coefficient for both MET sections. Table 4.11 presents a low exact level agreement but a high agreement for adjacent levels. This means that for almost 96% cases in Section I and 87% of cases in Section II, the center and cut scores classification agreed by exactly the same or by one level. This is further illustrated in Table 4.12

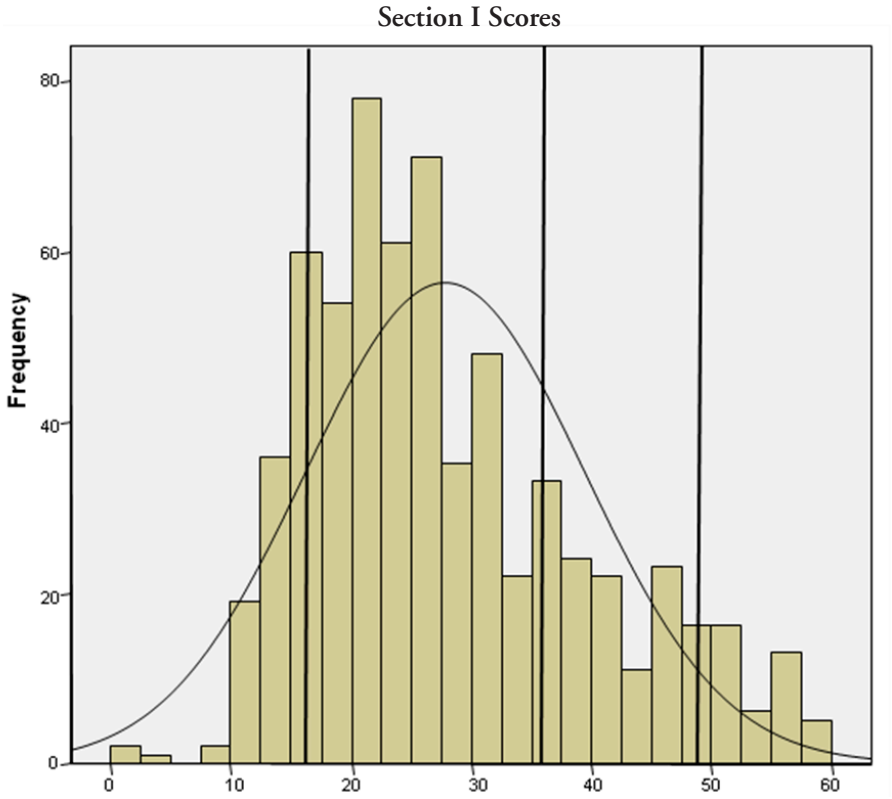

Figure 4.2 Section I Score Distribution and Cut Scores

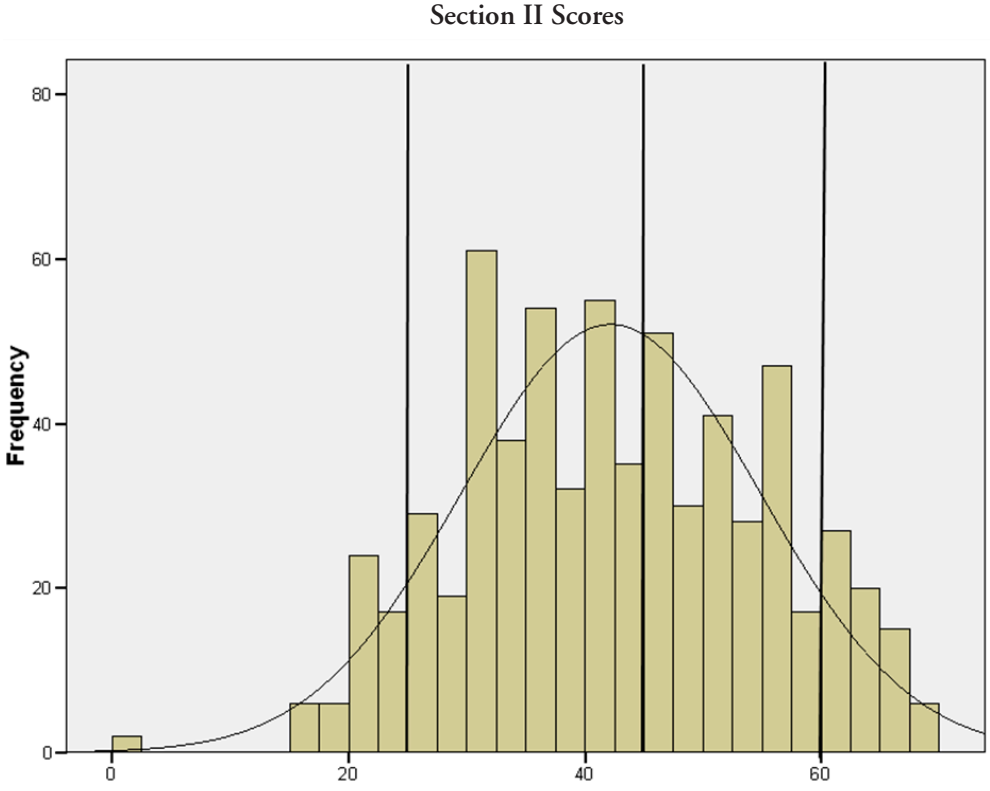

Figure 4.3 Section II Score Distribution and Cut Scores

and Table 4.13. Most of the students that the center classified as A2 received a B1 score in both sections. It should be explained, however, that the center probably chose the top performers from classes whose students were considered to be at A2 level. This was because the center was informed that the MET aimed primarily to test ability at levels higher than A2. Therefore, these students could have already reached B1 level when they took the test in July—that is, the end of the school year. Moreover, it was not possible to actually ask the teachers of these students to provide judgments of the students' proficiency level. Had this been done, a higher exact-level agreement could have been achieved. Despite these limitations of the data collection design, the high agreement within one level provides some support to the cut scores set in the study.

Table 4.10 Correlations of Level Classification Between the Test Center and the Cut Scores

	Section I cut scores	Section II cut scores
Centers	0.51	0.50

Table 4.11 Exact and Adjacent Level Agreement Between the Test Center and the Cut Scores

Agreement	Section I	Section II
Exact level	122 (40.40%)	92 (30.46%)
Within 1 level	290 (96.03%)	264 (87.42%)

Table 4.12 Cross-Tabulation of Level Classification Between the Test Center and the Cut Scores (Section I)

		Section I cut scores				Total
		2	3	4	5	
Center	2	33	118	7	1	159
	3	12	70	14	2	98
	4	1	21	9	2	33
	5	0	1	0	10	11
	6	0	0	0	1	1
Total		46	210	30	16	302

Table 4.13 Cross-Tabulation of Level Classification Between the Test Center and the Cut Scores (Section II)

		Section II cut scores				Total
		2	3	4	5	
Center	2	9	116	30	4	159
	3	2	55	37	4	98
	4	0	11	18	4	33
	5	0	0	1	10	11
	6	0	0	0	1	1
Total		11	182	86	23	302

4.7 THE JUDGES' FEEDBACK

Section 4 of this report has provided a variety of sources to support the validity of the recommended cut score. The section will conclude by looking at the results of the anonymous feedback questionnaire administered to the judges at the end of the last day. Eleven questionnaires were collected, with quantitative and quantitative data in the form of a four-point Likert scale and free comments respectively. The responses are summarized in Table 4.14.

Quantitative ratings were positive for the vast majority, offering further evidence of procedural validity. Some negative comments were made, as some judges felt they were in need of more familiarization with the CEFR and that there was disagreement with regard to the cut score. These comments serve as a reminder that standard setting is an inherently judgmental procedure, despite the use of empirical data to help judges with their task. However, it should be pointed out that the judges were not aware of the analyses conducted in this section with regard to cut score judgments (i.e., method and decision consistency and intra-judge and inter-judge consistency) and the exclusion of extreme ratings. Therefore, some of the judges might have felt that such extreme ratings affected the recommended cut scores too much. Moreover, the reasonableness of the cut scores suggested in this study was further examined by analyzing item responses from the administrations of the MET in 2009.

Table 4.14 Judges' Feedback Questionnaire Responses

	Not at all			Very	Missing
	1	2	3	4	
Question Were the two PowerPoint presentations on Day 1 informative with regard to standard setting and the CEFR linking?		1	2	7	1
Comments No comments written in the questionnaire					
Question Were the Familiarization tasks helpful in gaining a better understanding of the CEFR levels?		3	8		
Comments • We know we had to read before hand, but we didn't as much in depth as we could have so I feel I needed more. • We need more time to become more familiarized with the CEFR					
Question Were the instructions for Training with item difficulty clear?		3	8		
Comments • Perhaps we're not really familiar with some of the terminology regarding the understanding of statistics					

Table 4.14 Judges' Feedback Questionnaire Responses

	Not at all			Very	Missing
	1	2	3	4	
Question Were the instructions for the standard-setting method clear? *Comments* The coordinator presented the tasks in a very clear way			1	9	1
Question Was the use of statistics easy to follow during the discussion? *Comments* No comments written in the questionnaire		1	6	4	
Question How confident are you in the decisions you have made? *Comments* • I think once again, that with more exposure to the CEFR I might have been much more accurate • I believe more familiarization work with the CEFR should be done as a whole group • I think there are still more things to learn to become more confident in this aspect		1	8	2	
Question Were you given enough time to participate in the discussions? *Comments* No comments written in the questionnaire			1	10	
Question Did you have enough time to complete your individual tasks? *Comments* No comments written in the questionnaire			2	9	
Question Overall, did the coordinator guide the meeting effectively? *Comments* No comments written in the questionnaire				11	

Table 4.14 Judges' Feedback Questionnaire Response

Question

Please write here what you liked the most during the meeting.

Comments

- The description of the criteria for every single level
- The familiarization stage was good
- I liked the fact that we could compare perceptions and opinions on the way we linked the CEFR to the exam
- Very clear information
- To be able to understand better what sts are expected to do when they reach a specific level. Though it can be confusing to draw a line between one level and the other, it's interesting to see that there are some subtle differences among them
- We worked hard and went straight to the point interpret the CEFR standards
- The possibility to increase our understanding of the CEFR and to take it more seriously
- Getting to know more about the CEFR in the way every session was presented, thank you
- Everything was a new learning experience

Question

Please state here what you enjoyed the least during the meeting

Comments

- Adding all these numbers, we should be asked to bring a laptop to use an excel program to get the results faster and better
- None!
- I don't know if it is normal to have judges be really off, but I expected peers to kind of have similar understanding
- Nothing, thank you very much!
- Sometimes the sessions were too long!
- We need to keep this up
- I think many participants did not understand this exercise is intended to determine the cut scores based on CEFR for the MET. They assumed this had already been done and we were trying to get closer to it, when in fact, it does not exist yet. Also, some participants (most) don't have a clear understanding of the CEFR.
- Sometimes I tended to be very strict

5. CONCLUSION

This technical report has presented the setting of cut scores for the two sections of the MET on the level of the Common European Framework of Reference. Using Item Response Theory, ability values of the examinees at the raw data cut score were calculated and converted on the 0–80 range of scaled scores reported to MET examinees. After further item response analyses conducted in 2009 from live administrations, the range of MET scaled scores that correspond to the CEFR levels is presented in Table 5.1. For further details please consult the MET website (http://www.lsa.umich.edu/eli/testing/met).

Table 5.1 CEFR Level Equivalence of the MET Scaled Scores

	Section I	Section II
C1	64 and above	64 and above
B2	53–63	53–63
B1	40–52	40–52
A2	39 or below	39 or below

REFERENCES

Alderson, J. C. (1993). Judgements in language testing. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 46–57). Alexandria, VA: TESOL.

Alderson, J. C. (2005). Editorial. *Language Testing, 22*(3), 257–260.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington: American Council on Education.

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 (2003). *Manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment. Preliminary pilot version.* Strasbourg: Council of Europe.

Council of Europe (2009). Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment. A Manual. Retrieved 02/02/2009, from http://www.coe.int/T/DG4/Portfolio/documents/Manual%20Revision%20-%20proofread%20-%20FINAL.pdf

Davidson, F. (1996). *Principles of statistical data handling*. Thousand Oaks, CA: Sage Publications.

Figueras, N., North, B., Takala, S., Verhelst, N., & Van Avermaet, P. (2005). Relating examinations to the Common European Framework: a manual. *Language Testing*, *22*(3), 261–279.

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.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89–116). Mahwah, N.J.: Lawrence Erlbaum Associates.

Kaftandjieva, F. (2004). *Standard setting. Section B of the Reference Supplement to the preliminary version of the Manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*. Strasbourg: Council of Europe.

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.

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.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46.

Nichols, D. P. (2006). Choosing an intraclass correlation coefficient Retrieved 09/09/2006, from http://www.utexas.edu/its/rc/answers/spss/spss4.html.

Norcini, J. J., Lipner, R. S., Langdon, L. O., & Strecker, C. A. (1987). A comparison of three variations on a standard-setting method. *Journal of Educational Measurement*, *24*(1), 56–64.

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

North, B. (2000b). Linking language assessments: An example in a low stakes context. *System, 28*(4), 555–577.

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

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.

Takala, S. (Ed.). (2004). *Reference Supplement to the preliminary version of the Manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching and assessment*. Strasbourg: Council of Europe.

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.

APPENDIX 1

SAMPLE MATERIAL USED TO FAMILIARIZE JUDGES WITH THE CEFR LEVELS

Your Name: _____ **Your ID:** _____		
ID	**Statement**	**Level**
L01	Can catch the main point in short, clear, simple messages and announcements.	
L02	Can catch the main points in TV programmes on familiar topics when the delivery is relatively slow and clear.	
L03	Can easily follow complex interactions between third parties in group discussion and debate, even on abstract, complex unfamiliar topics.	
L04	Can extract specific information from poor quality, audibly distorted public announcements, e.g. in a station, sports stadium etc.	
L05	Can extrapolate the meaning of occasional unknown words from the context and deduce sentence meaning provided the topic discussed is familiar.	
L06	Can follow a lecture or talk within his/her own field, provided the subject matter is familiar and the presentation straightforward and clearly structured.	
L07	Can follow changes of topic of factual TV news items, and form an idea of the main content.	
L08	Can follow detailed directions.	
L09	Can follow extended speech even when it is not clearly structured and when relationships are only implied and not signalled explicitly.	
L10	Can follow films employing a considerable degree of slang and idiomatic usage.	

APPENDIX 2

SAMPLE MATERIAL USED TO TRAIN JUDGES WITH ITEM DIFFICULTY

Your Name: _____ **Your ID:** _____

Instructions

Respond to the items of the first reading set of the MET pilot Form B (pp. 12–13). Then, in the second column below, rank the items from easiest to the most difficult. Use the item number (81–96). After the end of the task, you will be given the ranking of the items based on their empirical difficulty (grey column). A discussion will follow where we will consider the following question: how accurate are judges when predicting item difficulty?

Item Ranking	Item #	Ranking Based on Empirical Difficulty
1 (easy)		
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15		
16 (diff)		

APPENDIX 3

SAMPLE MATERIAL USED TO COLLECT JUDGES' CUT SCORE ESTIMATES

Instructions

First take the MET Form A Listening Section. Then, think of 100 examinees, who are exactly on the border between two adjacent CEFR levels. Estimate how many of these 100 examinees will answer each item correctly and write the number (a whole number between 0 and 100) in the corresponding column. Then in the last two rows calculate the total of each column and then divide it by 100.

	Borders Between Levels		
Item Number	**A2 / B1**	**B1/B2**	**B2/C1**
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			
21			
22			
23			
24			
25			
26			
27			
28			
Total			
Total / 100			