MICHIGAN LANGUAGE ASSESSMENT

MET Go!

Linking the MET Go! and the Common European Framework of Reference

Technical Report

Sharon Pearce

Patrick McLain

Tony Clark

Susan Haines

CONTACT INFORMATION

All correspondence and mailings should be addressed to:

Michigan Language Assessment
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@michiganassessment.org
michiganassessment.org

MICHIGAN LANGUAGE ASSESSMENT
Cambridge Assessment English
UNIVERSITY OF MICHIGAN
11/2018

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1 Overview

This report summarizes the results of a standard setting study that was conducted to link scores on each section of the MET Go! to the proficiency levels of the Common European Framework of Reference. This study utilized the Council of Europe's (2009) manual supporting standard setting and Tannenbaum and Cho's (2014) article on critical factors to consider in standard setting as guidelines. This report documents the standard setting study and provides validity evidence to support its quality.

1.2 The MET Go!

MET Go! is a multi-level test of English language ability designed for beginner to intermediate level learners of middle and secondary school age. Developed and produced by Michigan Language Assessment, the test covers the four language skills (listening, reading, speaking, and writing), assessing learners' ability in each area and assisting them as they progress in their learning.

The listening and reading sections are comprised of three option selected response questions. The listening section has five parts: identifying people in a picture, short dialogue, listener-directed question, longer dialogue, and announcements, while the reading section has two parts: vocabulary/grammar and reading passages. All of the audio recordings in the listening section are played twice.

The speaking and writing sections are comprised of constructed response tasks. The writing section consists of three parts: writing a story about a set of pictures, writing to describe a personal experience, and writing to express a preference, while the speaking section consists of four parts: an unscored warm-up activity, a picture comparison task, a picture description task, and personal experience and preference questions.

1.3 Common European Framework of Reference

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

1.4 Standard Setting

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

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

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

2. METHODOLOGY

2.1 Panelists

One of the most important elements of a standard setting study is the panel of experts that make judgments on the location of the cut scores. Standard setting is often described as "fundamentally, a decision-making process" (Skorupski, 2012, p. 135), so it is important that the participants have good knowledge of the examination in question, the test-taking population, and the performance level descriptors (Mills, Melican, & Ahluwalia, 1991; Papageorgiou, 2010).

This standard setting study utilized four separate panels, each evaluating one of the four sections of the MET Go! (listening, reading, speaking, and writing). The panelists were all recruited from Cambridge Assessment English staff and its network of external consultants, and were assigned to each panel based on their area of expertise. The listening and reading panels each consisted of twelve panelists, while the speaking and writing panels consisted of eleven. Table 2.1.1 summarizes the average years of experience for each panel in a number of relevant areas. Overall, it shows that the panelists selected for each of the study panels provided a diverse representation of experienced professionals in the field of ESL/EFL.

2.2 Standard Setting Method

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

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

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

For the MET Go! standard setting study, the panels applied the Angoff method to the listening and reading sections and the bookmark method to the speaking and writing sections in order to make three cut scores judgements (A1, A2, and B1) for each section of the exam. The Angoff and bookmark methods were selected for this study because they are well suited to setting cut scores on selected response and constructed response tests, respectively.

For the selected response sections, a live MET Go! form was used for the judgement round. To make their judgments the panelists were asked to consider 100 just-qualified candidates at each CEFR level and state for each item how many of the just-qualified candidates would answer it correctly. Panelists were asked to go through the test separately for each target CEFR level to make their just-qualified judgements so that they would only have to focus on one just-qualified definition at a time.

Table 2.1.1: Average Years of Experience for Each Panel

Experience	Listening	Reading	Speaking	Writing
ESL/EFL Teaching	18.58	18.67	16.27	13.91
Assessment/Test Development	10.25	9.75	10.64	8.09
Item Writing	6.42	6.00	2.64	3.09

For the constructed response sections, ordered item booklets[1] were created by selecting test taker performances for each possible score point on the rating scales and ordering them from lowest to highest. Each performance had been scored by at least two certified raters who worked to build a consensus on each performance's score. To make their cut score judgments, the panelists went through the ordered item booklets and placed their bookmarks at the first performance that they felt could have been produced by a just-qualified A1-, A2-, and B1-level candidate.

2.3 Meeting Procedures

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

Prior to the standard setting meetings, the panelists were required to complete several pre-study activities to begin familiarizing themselves with the MET Go! and the CEFR. After completing a brief background questionnaire, the panelists were also asked to complete a pre-study CEFR quiz to assess their understanding of the CEFR prior to the standard setting meetings. This quiz required panelists to assign CEFR levels to 18 descriptors selected from several scales related to the test section being linked. Once the quiz was completed, the panelists were asked to familiarize themselves with the MET Go! by reading information on the Michigan Language Assessment website. They were also asked to complete a pre-study activity to familiarize themselves with the CEFR levels relevant to the study and the concept of the just-qualified test taker. The activity asked the panelists to review the CEFR global scale (Council of Europe, 2001, p. 24), self-assessment gird (Council of Europe, 2001, p. 26–27), and several scales relevant to each panel (see appendix A for the list of scales selected for each panel), and then describe their initial impressions of the characteristics of an average and a just-qualified A1-, A2-, and B1-level candidate. See Appendix B for an example of the pre-study activity questions, which were taken (with some modification), from the Tannenbaum and Wylie (2008) standard setting report.

Each standard setting meeting began with a brief introduction to the standard setting procedure and the goals of the study, before moving onto the familiarization activities. To familiarize the panelists with the CEFR levels and descriptors, each panel participated in a familiarization activity that utilized descriptors from CEFR scales related to the panel's test section (see Appendix A). Panelists were given a set of decontextualized descriptors from these scales and asked to sort them and individually assign CEFR levels to each descriptor. Because these scales were not discussed prior to the activity, panelists needed to use their knowledge and understanding of the CEFR to help them complete the activity. The results of this activity were then discussed in detail as a group to ensure that the panelists understood the descriptors for each CEFR level. While this sorting activity can be rather challenging due to the decontextualization of the descriptors, it helped to encourage panelist familiarization with the CEFR by forcing them to fully read and deeply consider the language of each individual descriptor. The sorting activity was then followed by a detailed discussion of the pre-study activity. This discussion focused on the concept of the just-qualified test taker and its importance in standard setting. With guidance from the meeting facilitators, the panelists worked to come up with a shared understanding of the just-qualified A1, A2, and B1 test takers. These definitions were written on a whiteboard in the room, and once finalized were left up as a reference for the remainder of the standard setting meeting.

The familiarization activities were followed by a training activity that taught the panelists how to make cut score judgements using the Angoff (listening and reading panels) or bookmark (speaking and writing panels) method. The meeting facilitators demonstrated how to access the necessary materials (e.g. test booklet, ordered item booklet) via Google drive and how to utilize the data collection spreadsheet. After

[1] Note that the ordered item booklets were actually digital folders that contained ordered audio or pdf files of speaking or writing performances, not a physical booklet. In practice these digital folder are used in the same way as a physical booklet

demonstrating how to make the cut score judgments, each panel discussed the procedures to address any questions or concerns. The listening and reading panels were also given an opportunity to practice making cut score judgments with some sample test items since the Angoff method can be more difficult for panelists to learn than the bookmark method. Once the panelists were satisfied with their understanding of the standard setting method they were given a pre-judgment survey to assess their understanding of the procedures and their willingness to proceed with the judgment activity.

For the judgment activity, each panel followed the same procedures that they practiced during the training activity to make their cut score judgments for the just-qualified A1, A2, and B1 level test takers. The meeting facilitators again emphasized the importance of thinking about the just-qualified candidate at each level when making their decisions. Each panel was provided with the appropriate judgment materials for the test section: a live MET Go! listening or reading section booklet for the listening and reading panels, and an ordered item booklet of writing or speaking performances at each possible score point for the speaking and writing panels. Throughout the judgement process panelists had access to their personal notes, the CEFR scales that had been discussed during the familiarization activities, and the shared just-qualified candidate definitions written on the whiteboard in the room.

The judgment activity consisted of two judgment rounds where panelists marked their decisions on spreadsheets. Both judgment rounds were followed by a group discussion of the results. The discussion of the first judgment round allowed panelists to review the items and materials and discuss the reasoning behind their cut score decisions. The panelists reviewed several test items (listening and reading panels) and test taker performances (writing and speaking panels) as a group so that they could discuss the factors that influenced their decisions. The listening and reading panels were also provided with IRT difficulty statistics for each item to consider during the discussions. For the second judgment round the panelists were instructed to perform the judgment activity again, taking into account the discussions of the first judgment round, and, if they felt it was necessary, make adjustments to their cut score decisions. The discussion of the second judgment round focused on finalizing the panel's cut score recommendations. Once the cut score recommendations were finalized, the panelists were given a post-judgment survey to collect their opinions the quality of the meeting and their confidence in the recommended cut scores, as well as a post-study CEFR quiz to assess whether their knowledge of the CEFR descriptors changed after the meeting.

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

3. RESULTS

3.1 Specification

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

It is justifiable to link the MET Go! to the CEFR because each section of the exam was specifically developed to assess test takers' English listening, reading, speaking, and writing abilities at the A1-B1 levels of the CEFR. Both the CEFR (Council of Europe, 2001), and the CEFR companion volume (Council of Europe, 2018) were used by the test development teams throughout the development process as references against which to define the test constructs reflect in the MET Go!. Each section of the exam utilizes several different item and task types in order to capture performance information on test takers across the target ability range. More detailed information on the development of each section of the MET Go! can be found in the test development reports (see http://michiganassessment.org/about-us/research/).

3.2 Familiarization

This section summarizes the results of the familiarization activities conducted during each standard setting meeting. These activities help the panelists establish a greater familiarity and understanding of the CEFR levels relevant to the study, which in turn helps to improve the quality of their judgments and the validity of their final cut score recommendations.

Tables 3.2.1-3.2.4 summarize the familiarization activity results of the individual panelists for each panel. They present the number and percentage of descriptors placed at the correct CEFR level, the Spearman correlation (ρ) between CEFR levels assigned by the panelists and the correct CEFR levels, and a number representing the average CEFR level assigned for each panelist. The correlation coefficient shows the degree to which the panelists understand the progression of the CEFR levels and should be interpreted in conjunction with the number and percentage of descriptors correct to understand the panelists' performance on the familiarization tasks. The average assigned CEFR level for each panelist was calculated by transforming their assigned CEFR levels to numbers (Pre-A1 = 1, A1 = 2, A2 = 3, B1 = 4, B2 = 5) and taking the average. The panelists' averages can be compared with the average level of the descriptors to assess their overall severity or leniency. Panelists with average assigned CEFR levels higher than the actual average were generally more lenient, while panelists with average assigned CEFR levels lower than the actual averages were generally more severe.

The results summarized in Tables 3.2.1-3.2.4 suggest that each panelists had a good understanding of the CEFR levels relevant to the study. The large average percentage of descriptors assigned to the correct CEFR level (57.96%-73.22%) and the high average correlation coefficients (0.855-0.929) for each panel provide evidence that the panelists were typically able

Table 3.2.1: Listening Panel Familiarization Activity Results (37 Descriptors, 3.27 Average CEFR Level)

Measure	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	Average
# Correct	25	13	25	20	22	28	27	21	21	22	25	19	22.33
% Correct	67.57	35.14	67.57	54.05	59.46	75.68	72.97	56.76	56.76	59.46	67.57	51.35	60.36
Correlation (ρ)	0.926	0.871	0.908	0.885	0.810	0.930	0.912	0.875	0.886	0.847	0.897	0.919	0.889
Average	3.00	2.68	3.03	2.89	3.14	3.24	3.19	3.27	2.84	3.38	3.08	2.95	3.06

Table 3.2.2: Reading Panel Familiarization Activity Results (38 Descriptors, 3.32 Average CEFR Level)

Measure	R1	R2*	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	Average
# Correct	28	16	20	21	21	18	22	27	20	18	25	20	21.33
% Correct	73.68	64.00	52.63	55.26	55.26	47.37	57.89	71.05	52.63	47.37	65.79	52.63	57.96
Correlation (ρ)	0.917	0.754	0.861	0.901	0.798	0.869	0.841	0.910	0.929	0.767	0.893	0.816	0.855
Average	3.53	3.44	3.16	2.95	3.47	3.11	3.08	3.13	2.84	3.87	3.13	3.34	3.25

** Several of R2's responses were not recorded due to an issue with response collection, so R2's summary is based on the 25 responses that were recorded.*

Table 3.2.3: Speaking Panel Familiarization Activity Results (37 Descriptors, 3.19 Average CEFR Level)

Measure	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	Average
# Correct	35	25	30	29	24	25	25	31	27	23	24	27.09
% Correct	94.59	67.57	81.08	78.38	64.86	67.57	67.57	83.78	72.97	62.16	64.86	73.22
Correlation (ρ)	0.994	0.922	0.949	0.936	0.898	0.910	0.904	0.965	0.932	0.897	0.911	0.929
Average	3.24	3.03	3.22	3.35	3.14	3.24	2.89	3.30	3.03	2.86	3.22	3.14

Table 3.2.4: Writing Panel Familiarization Activity Results (37 Descriptors, 3.51 Average CEFR Level)

Measure	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	Average
# Correct	21	22	17	31	22	23	30	25	24	15	31	23.73
% Correct	56.76	59.46	45.95	83.78	59.46	62.16	81.08	67.57	64.86	40.54	83.78	64.13
Correlation (ρ)	0.795	0.779	0.796	0.942	0.828	0.876	0.947	0.839	0.824	0.832	0.956	0.856
Average	3.41	3.54	3.19	3.57	3.24	3.35	3.59	3.54	3.43	3.08	3.41	3.40

to correctly identify the level of the descriptors and that they understood the progression of language across the different CEFR levels. Additionally, the average assigned CEFR levels indicate that, while each panel tended to be slightly severe as a group, there was still a good amount of variation in the leniency/severity of the individual panelists.

When assessing familiarity of the panelists with the CEFR it is also important to consider the level of agreement and consistency of the panel as a whole. Table 3.2.5 presents three measures of internal consistency for each panel's familiarization activities: Cronbach's alpha (α), the intraclass correlation coefficient (ICC), and Kendal's coefficient of concordance (W). These three indices are among the most frequently used measures of internal consistency (Kaftandjieva, 2010, p. 96). Cronbach's alpha measures internal consistency by estimating the proportion of variance due to common factors in the items (Davies et al., 1999, p. 39), the ICC measures internal consistency by taking into account both between- and within-rater variance (Davies et al., 1999, p. 89), and Kendall's W is a nonparametric measure of internal consistency that measures the level of agreement between three or more raters that rank the same group of items (Davies et al., 1999, p. 100). These three indices range from 0 to 1, with a value of 1 indicating complete agreement among panelists. Table 3.2.5 shows that all three of these indices were very high, with Cronbach's alpha and ICCE values very close to 1, which suggests that there was a very high level of agreement and consistency in the panelists understanding of the CEFR for each panel.

Table 3.2.5: Panel Agreement and Consistency for Familiarization Activities

Panel	α	ICC*	W
Listening	0.985	0.983	0.819
Reading**	0.980	0.974	0.789
Speaking	0.988	0.987	0.859
Writing	0.975	0.973	0.754

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

*** Several of R2's responses were not recorded due to an issue with response collection, so R2's responses are excluded from the Reading panel calculations.*

Finally, a measure of the overall effectiveness of the familiarization tasks can be obtained by analyzing the results of the pre- and post-study CEFR quizzes. As discussed in Section 2.3, each panelists was asked to complete two short (18 item) CEFR quizzes prior to and immediately following the standard setting meeting in order to assess their initial understanding of the CEFR and determine whether their understanding of the CEFR had improved. Figure 1 presents scatterplots that summarize the individual results of the pre- and post-study quiz for each panel. Examination of the panelists' pre- and post-study quiz scores relative to the identity line (represented by the dashed line in Figure 1) reveals that while most panelists performed similarly on the pre- and post-study quizzes, many of the panelists did perform slightly better on the post-study quiz. Table 3.2.6 summarizes the average pre- and post-study quiz results for each panel, as well as the results of paired t-tests. It shows that, on average, while the panelists' scores did improve for each panel after the standard setting meeting, this difference was only statistically significant for the writing panel.

When evaluating the familiarization activities it is important to note that they are meant to be learning activities that expose the panelists to CEFR descriptors relevant to the study and ensure that they all have an accurate understanding of each CEFR level. Therefore, some inaccuracies and inconsistencies in panelists' responses during the activity are expected. The descriptor statements were thoroughly discussed after each familiarization activity to address any questions and ensure that the panelists understood the correct levels of each descriptor. Furthermore, this familiarization activity was followed by a discussion of what characteristics defined the just-qualified candidate at each relevant CEFR level to further enhance the panelists' ability to make high quality judgements.

Overall, the analysis of the familiarization activities suggests that each panel had a good understanding of the CEFR. Assigning CEFR levels to individual descriptors is a challenging task, but the results summarized above show that the individual panelists had a good understanding of the CEFR levels, that each panel had a high level of agreement and consistency, and that the familiarization activities and discussions were successful in helping to improve upon many of the panelists' understanding of the CEFR. The comments made by panelists throughout the discussion of the familiarization activities, the responses to the pre- and post-judgment surveys (see Section 4.1), and the low variability of the judgment task (see Section 3.3) provide evidence to further support the panelists' understanding of the relevant CEFR levels.

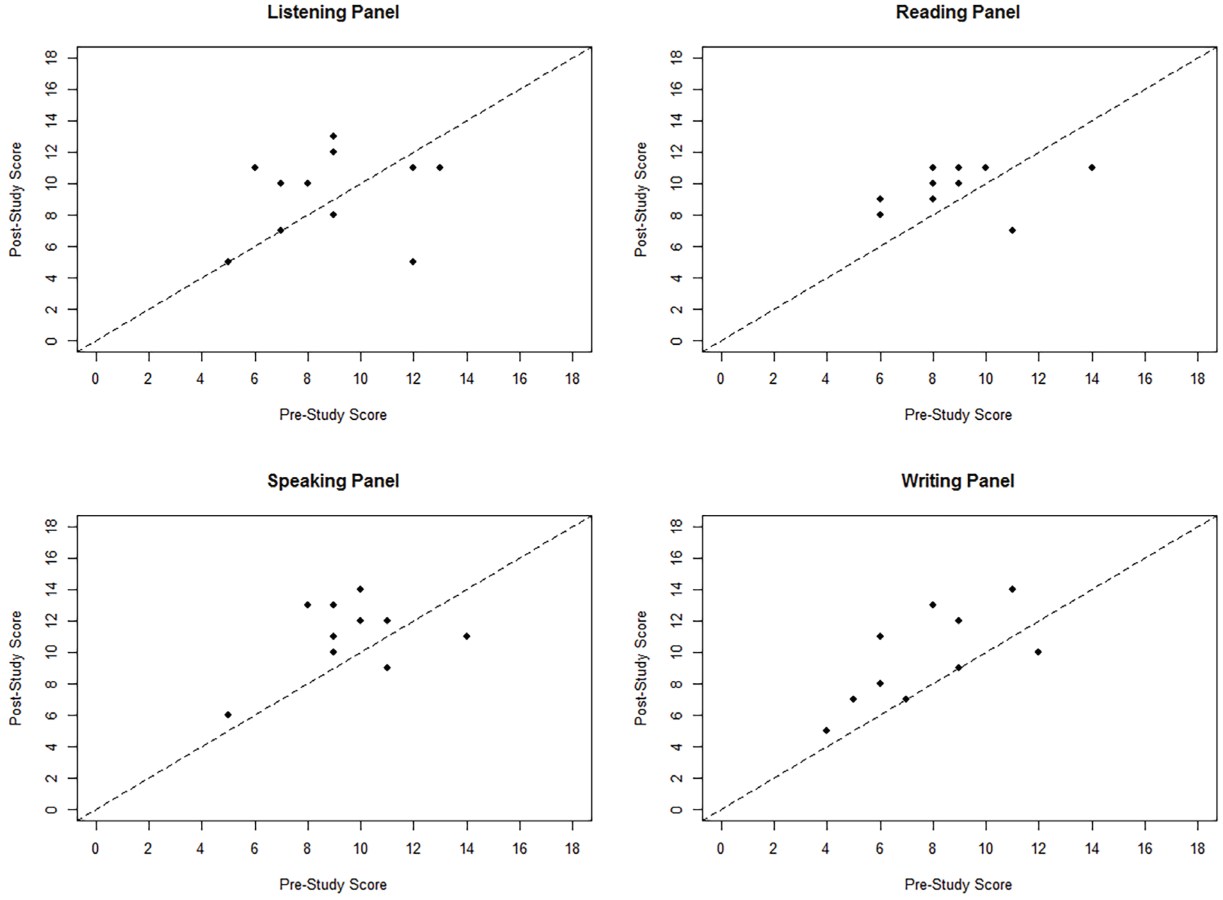

Figure 1: Summary of Pre- and Post-Study CEFR Quiz Results

Table 3.2.6: Summary of Pre- and Post-Study CEFR Quiz Results

Panel	Average Correct (out of 18)		Paired T-Test Results		
	Pre-Study	Post-Study	t	df	p-value
Listening	8.91	9.27	0.79	11	0.449
Reading	9.27	9.91	1.33	11	0.210
Speaking	9.70	10.80	1.99	10	0.075
Writing*	7.70	9.60	2.69	9	0.025

** W5's results are not included in this analysis since they did not complete the post-judgement quiz.*

3.3 Judgment

This section summarizes the results of the judgement activities conducted during each standard setting meeting. For these activities, the panelists' carefully evaluated each test item or performance and utilized the Angoff (listening and reading panels) and bookmark (speaking and writing panels) methods to make their judgments and arrive at their individual cut score recommendations.

Table 3.3.1-3.3.4 summarize the results of the judgement activities for each panel. The tables present the panelists' individual cut score recommendations and summary statistics for each panel's recommendations from each judgement round. They show that the panelists' cut score recommendations were all quite similar, and that there was very little variation in the panelists' individual cut score recommendations.

Table 3.3.1: Listening Panel Cut Score Judgments

Panelist ID	Judgement Round 1			Judgement Round 2		
	A1	A2	B1	A1	A2	B1
L1	5.65	14.10	19.10	6.20	14.25	20.15
L2	7.90	14.25	20.08	7.42	14.70	20.98
L3	7.42	15.38	22.36	7.32	15.43	22.29
L4	6.85	21.75	28.20	7.80	18.40	23.70
L5	12.36	20.14	25.90	13.18	19.43	24.19
L6	8.65	14.25	21.60	8.25	14.90	21.65
L7	5.70	12.50	21.21	5.36	12.57	21.99
L8	8.60	14.90	20.95	8.50	14.75	21.00
L9	1.75	14.20	23.50	8.10	16.50	24.30
L10	3.80	9.45	16.30	6.90	13.95	19.90
L11	9.15	16.28	23.26	8.55	16.05	23.26
L12	6.95	17.15	21.15	7.15	16.90	21.25
Average	7.07	15.36	21.97	7.89	15.65	22.06
SD	2.71	3.25	3.09	1.91	1.93	1.51
Median	7.19	14.58	21.41	7.61	15.17	21.82
Minimum	1.75	9.45	16.30	5.36	12.57	19.90
Maximum	12.36	21.75	28.20	13.18	19.43	24.30

Table 3.3.2: Reading Panel Cut Score Judgments

Panelist ID	Judgement Round 1			Judgement Round 2		
	A1	A2	B1	A1	A2	B1
R1	6.25	12.95	20.30	6.95	13.95	21.10
R2	3.85	10.40	17.70	5.55	12.35	19.50
R3	4.80	14.55	23.05	5.10	14.85	23.08
R4	5.55	14.75	22.50	5.55	14.65	22.25
R5	9.40	15.00	21.75	8.35	14.55	22.05
R6	5.55	17.70	24.50	4.80	16.65	23.80
R7	8.95	16.65	22.34	7.30	15.95	22.90
R8	8.40	15.45	23.76	7.77	15.02	23.21
R9	6.20	14.65	23.90	6.25	14.30	23.10
R10	8.20	14.25	21.82	8.32	14.25	21.47
R11	8.25	13.15	19.05	8.70	13.60	19.60
R12	11.50	18.30	22.65	9.40	16.35	21.15
Average	7.24	14.82	21.94	7.00	14.71	21.93
SD	2.22	2.14	2.02	1.54	1.20	1.41
Median	7.23	14.70	22.42	7.13	14.60	22.15
Minimum	3.85	10.40	17.70	4.80	12.35	19.50
Maximum	11.50	18.30	24.50	9.40	16.65	23.80

Table 3.3.3: Speaking Panel Cut Score Judgments

Panelist ID	Judgement Round 1			Judgement Round 2		
	A1	A2	B1	A1	A2	B1
S1	4	11	16	5	9	13
S2	7	9	14	5	9	13
S3	5	8	13	5	9	13
S4	5	9	13	5	9	13
S5	7	9	15	5	9	13
S6	8	10	13	5	9	13
S7	7	13	15	5	10	15
S8	8	12	16	5	9	15
S9	5	9	14	5	9	13
S10	5	13	15	5	8	15
S11	7	12	15	5	9	13
Average	6.18	10.45	14.45	5.00	9.00	13.55
SD	1.40	1.81	1.13	0.00	0.45	0.93
Median	7	10	15	5	9	13
Minimum	4	8	13	5	8	13
Maximum	8	13	16	5	10	15

Table 3.3.4: Writing Panel Cut Score Judgments

Panelist ID	Judgement Round 1			Judgement Round 2			Judgement Round 3
	A1	A2	B1	A1	A2	B1	A2*
W1	8	14	22	7	11	23	15
W2	8	12	24	8	12	24	15
W3	8	11	17	8	11	17	11
W4	7	15	24	7	12	24	15
W5	8	18	24	8	15	24	15
W6	8	18	27	8	17	24	17
W7	11	19	24	8	17	24	17
W8	7	12	23	8	12	24	16
W9	12	23	27	11	17	24	17
W10	10	16	25	9	17	24	17
W11	5	8	24	7	15	24	15
Average	8.36	15.09	23.73	8.09	14.18	23.27	15.45
SD	1.96	4.28	2.69	1.14	2.60	2.10	1.75
Median	8	15	24	8	15	24	15
Minimum	5	8	17	7	11	17	11
Maximum	12	23	27	11	17	24	17

** A third judgement round was conducted for the A2 cut that ignored writing script 12, since several panelists felt that it was out of order compared to the rest of the writing scripts.*

4. VALIDITY EVIDENCE

4.1 Procedural Validity

The documentation of the standard setting study throughout this report provides procedural validity evidence to support the quality of the standard setting meetings and the cut score recommendations. This section provides additional procedural validity evidence by summarizing the panelists' responses to pre- and post-judgment surveys that were given during the standard setting meetings. The pre-judgment survey focused on the panelists' understanding of the familiarization and training activities, while the post-judgment survey focused on the panelists' understanding of the judgment rounds and their confidence in the recommended cut scores. Both surveys used a four-point Likert scale (1 – strongly disagree to 4 – strongly agree) to collect most of this information. Tables 4.1.1 and 4.1.2 present the statements and summarize the results for the pre- and post-judgment surveys, respectively. In addition to the statements listed in the tables, the pre-judgment survey asked the panelists to indicate if they were ready to proceed to the judgment task (yes or no), and the post-judgment survey asked the panelists to indicate their opinion of the recommended cut scores (too low, about right, or too high).

These tables show that the panelists generally responded favorably to the survey statements across all panels. The majority of panelists indicated that they understood the familiarization, training, and judgment activities and expressed confidence in their decisions and indicated that they had enough time to complete their tasks and participate in group discussions. On the pre-judgment survey, all of the panelists indicated that they felt ready to continue to the judgment activity. On the post-judgment survey, 12 listening panelists, 11 reading panelists, 10 speaking panelists, and 10 writing panelists indicated that the cut score recommendations were about right, and 1 speaking panelist and 1 writing panelist indicated that the cut score recommendations were too low. Out of the 45 panelists that participated in these linking activities, only 5 disagreed with any of the survey statements.

Table 4.1.1: Summary of Pre-Judgment Survey Results

Question	Statement	Listening				Reading				Speaking				Writing			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	The familiarization activities helped me to understand the CEFR levels	-	-	5	7	-	-	7	5	-	-	7	4	-	-	7	4
2	The training activity helped me to understand the judgment process	-	-	2	10	-	-	2	10	-	-	2	9	-	-	5	6
3	I had enough time to complete my individual tasks	-	-	2	10	-	1	1	10	-	-	1	10	-	-	4	7
4	I had enough time to participate in the discussions	-	-	2	10	-	-	2	10	-	-	1	10	-	-	3	8

Table 4.1.2: Summary of Post-Judgement Survey Results

Question	Statement	Listening				Reading*				Speaking				Writing			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	The familiarization activities helped me to understand the CEFR levels	-	-	5	7	-	-	7	5	-	-	7	4	-	-	7	4
2	The training activity helped me to understand the judgment process	-	-	2	10	-	-	2	10	-	-	2	9	-	-	5	6
3	I had enough time to complete my individual tasks	-	-	2	10	-	1	1	10	-	-	1	10	-	-	4	7
4	I had enough time to participate in the discussions	-	-	2	10	-	-	2	10	-	-	1	10	-	-	3	8
5	I had enough time to complete my individual tasks	-	-	3	9	-	-	1	10	-	-	3	8	-	-	3	8
6	I had enough time to participate in the discussions	-	-	2	10	-	-	2	9	-	-	1	10	-	-	5	6
7	I am confident in the decisions I have made	-	-	4	8	-	-	7	4	-	1	4	6	-	2	7	2

** One reading panelist did not fill out the post-judgement survey, so the table only summarizes the post-judgment survey responses of 11 panelists.*

One of the reading panelists disagreed with pre-judgement statement 3, indicating that she did not feel as though she had enough time to complete her individual tasks. However, she expressed agreement with all the other pre-judgement statements, which suggests that her initial dissatisfaction with the amount of time available to complete the tasks did not negatively impact her understanding of the CEFR or the judgement process. One of the listening panelists disagreed with post-judgment statement 3, indicating that she did not understand the instructions for each judgment round. However this panelist also commented that her misunderstand stemmed from focusing too much on how to play the audio, and that the meeting facilitators were able to sort out her misunderstandings during the meeting. Finally, one of the speaking panelists and two of the writing panelists disagreed with post-judgement statement 7, indicating that they were not confident in the decisions they made. One of the speaking panelists indicated that their lack of confidence was due to fatigue they were feeling at the end of the day, but still indicated that the final recommended cut scores seemed about right. The other speaking panelist and the writing panelist also indicated that they felt the recommended cut scores were too low.

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

4.2 Internal Validity

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

SEM estimates for each section were obtained using MET Go! pilot test data. The panelists' judgments for the listening, reading, speaking, and writing panels were not originally made using MET Go! scaled scores, so their raw cut score recommendations needed to be transformed onto the appropriate scale before we could calculate the SEj values for comparison. This was done by rounding the panelists' cut score recommendations to the nearest whole number and applying the appropriate raw-to-scale conversion table. Table 4.2.1 presents the SEj values for each panel's cut scores, as well as the SEM estimates for each test section based on available MET Go! pilot data. It shows that the SEj values are much less than half of each section's SEM value for each panels, which suggests that the panel's cut score recommendations are dependable and that they would likely be replicated in another standard setting study.

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

The summary statistics and reliability estimates from above were also used here, in conjunction with the formula for calculating standard z scores and the tables from Subkoviak (1988), to estimate the agreement and kappa coefficients for each cut score. Table 4.2.2 summarizes these estimates for each section's overall cut scores. It should be noted that for high-stakes exams reliability estimates of 0.80 and above are expected and

Table 4.2.1: Standard Error of Judgment for Each Panel

Panel	SEM	SEM/2	SEj		
			A1	A2	B1
Listening	4.81	2.41	1.093	1.067	1.086
Reading	3.18	1.59	0.747	0.633	0.838
Speaking	3.14	1.57	0.000	0.423	0.704
Writing	4.17	2.09	0.513	0.877	1.088

Table 4.2.2: Agreement Coefficient (p_0) and Kappa (κ) for Panel Cut Scores

Cut Score	Listening		Reading		Speaking		Writing	
	p_0	κ	p_0	κ	p_0	κ	p_0	κ
B1	0.86	0.71	0.87	0.71	0.87	0.70	0.86	0.71
A2	0.89	0.70	0.90	0.68	0.98	0.58	0.87	0.71
A1	0.95	0.63	0.98	0.58	0.98	0.58	0.91	0.68

acceptable. Thus, based on Subkoviak's (1988) tables, we should expect agreement coefficients greater than or equal to 0.80, and kappa coefficients between 0.45 and 0.71. Table 4.2.2 shows that the agreement and kappa coefficients are generally quite high (p_0 ≥0.86, κ ≥0.58) for each cut score. These strong agreement coefficients suggest that test takers would likely be consistently classified into the same CEFR level if they were to take the exam multiple times, and the strong kappa coefficients of these cut scores suggest that the reliability of the test scores has a good contribution to the overall classification consistency.

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

4.3 External Validity

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

Even so, this report attempts to provide some external validity evidence by exploring the reasonableness of the recommended cut scores. This was done by applying the recommended cut scores to MET Go! pilot test results and examining the resulting CEFR distributions. Table 4.3.1 presents the CEFR distributions for each section. It shows that the CEFR distributions for the listening, reading, and writing sections are all quite similar, though the reading section distribution suggests that the pilot test takers may have been less proficient in their reading ability. The CEFR distribution of the speaking section differs greatly from those of the other three sections, but this is because the test takers that opted to participate in the pilot speaking test tended to be more proficient users of English than those who opted not to participate. Overall, these CEFR distributions are in line with our expectations for the pilot population, which helps to provide some external validity evidence for the MET Go! cut scores.

Table 4.3.1: Distribution of MET Go! Pilot Test Taker by CEFR Level

Section	N	Below A1	A1	A2	B1
Listening	670	13.73	13.58	20.15	52.54
Reading	670	0.15	23.43	38.81	37.61
Speaking	180	0.56	5.00	23.33	71.11
Writing	563	23.80	9.77	16.16	50.27

5. CONCLUSION

This report has provided a detailed summary of the standard setting study conducted to link MET Go! scores to the CEFR. It has documented both the procedures and results of the study, including the standard setting meetings, and has provided procedural, internal, and external validity evidence to support the quality of the cut score recommendations. The raw cut score recommendations made by the panelists for each panel were used to inform the final cut score placement, which were then scaled onto the MET Go! score reporting scale. Table 5.1 summarizes the CEFR score bands for MET Go! scaled scores.

Table 5.1: MET Go! CEFR Score Bands

CEFR Level	Scaled Score Range
B1	40 – 52
A2	27 – 39
A1	14 – 26
Below A1	0 – 13

6. REFERENCES

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

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

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

Council of Europe. (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge, UK: 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 https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680667a2d

Council of Europe. (2018). *Common European Framework of Reference for Languages: learning, teaching, assessment. Companion Volume with New Descriptors.* Retrieved from https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Appendix A: CEFR Scales Used for each Familiarization Activity

Table A.1: CEFR Scales Used in Familiarization Activities for Each Panel

CEFR Scale Title	Page Number (Council of Europe, 2018)	Study Panel
Overall Listening Comprehension	55	Listening
Understanding Conversation Between Other Speakers	56	Listening
Listening to Announcements and Instructions	58	Listening
Overall Reading Comprehension	60	Reading
Reading Correspondence	61	Reading
Reading for Information and Argument	63	Reading
Creative Writing	76	Writing
Written Reports and Essays	77	Writing
Correspondence	94	Writing
Vocabulary Range	132	Writing
Grammatical Accuracy	133	Writing
Thematic Development	141	Writing
Coherence and Cohesion	142	Writing
Overall Spoken Production	69	Speaking
Sustained Monologue: Describing Experience	70	Speaking
Sustained Monologue: Giving Information	71	Speaking
General Linguistic Range	131	Speaking
Spoken Fluency	144	Speaking

Appendix B: Example Pre-study Activity (Writing Panel)

Your name (first and last)

____

Directions: Clink the links below and read through the CEFR scales. Then complete the short exercise.

1. Global Scale and Self-Assessment Grid
2. Writing[2] Scales

1. Based on the information in the CEFR scales, please describe what you perceive are the key characteristics of an **average A1 writer**[3].

   ____

   ____

2. Based on the information in the CEFR scales, please describe what you perceive are the key characteristics of a **just-qualified (i.e., minimally competent) A1 writer**.

   ____

   ____

3. Based on the information in the CEFR scales, please describe what you perceive are the key characteristics of an **average A2 writer**.

   ____

   ____

4. Based on the information in the CEFR scales, please describe what you perceive are the key characteristics of a **just-qualified (i.e., minimally competent) A2 writer**.

   ____

   ____

5. Based on the information in the CEFR scales, please describe what you perceive are the key characteristics of an **average B1 writer**.

   ____

   ____

6. Based on the information in the CEFR scales, please describe what you perceive are the key characteristics of a **just-qualified (i.e., minimally competent) B1 writer**.

   ____

   ____

[2] The writing scales were substituted with listening, reading, and speaking scales in the other panels' activities.

[3] The term "writer" was substituted with "listener", "reader", and "speaker" in the other panels' activities.