ITSC International Testing Services Center G-TELP™ General Tests of English Language Proficiency

Linking an English Language Test (G-TELP) to the CEFR: A Comparison of Modified Angoff and Bookmark Methods

2023

Prepared by
Minjung Kim, PhD

Conference
Asia TEFL

Date
Aug 20, 2023

Linking an English Language Test(G-TELP) to the CEFR: A Comparison of Modified Angoff and Bookmark Methods

This study aimed to link a large-scale English proficiency test to the Common European Framework of Reference for Languages (CEFR). CEFR was introduced and applied systematically to consistently evaluate the contents and methods of a large-scale English proficiency test, G-TELP. Standard-setting for the CEFR was conducted by a panel of 10 members from various countries. The panel members judged cut scores in the Listening, Grammar Vocabulary, and Reading sections using the Modified Angoff and Bookmark methods. By comparing the modified Angoff method and the bookmark method, the appropriate procedures were used to secure its validity. The methods of assigning competency description items were intuitive, qualitative, and quantitative. Based on these judgments, cut scores that divided test takers into the six levels of the CEFR were derived, and the validity of the cut scores was evaluated using internal, external, and procedural criteria.

The analysis showed how the judgment's standard deviation changes across rounds; the cut scores derived from the two methods were similar or not, and almost all the panel members expressed confidence in the final cut scores on the post-standard-setting survey. Furthermore, most panel members responded that they found the preparation materials, the standard-setting process, and the facilitators' explanations to be clear and helpful when making judgments. CEFR showed whether the English language test emphasized sociolinguistic knowledge and language strategies or not grammar and vocabulary-centered teaching methods.

Keyword: Standard-setting, CEFR, Modified Angoff, G-telp Level 2, Bookmark method

I. Purpose of research

The calculation of achievement standard information on the extent to which test takers have achieved achievement standards for each area in the test is used as basic data for establishing test policies and plans. For example, in the case of the national level evaluation of academic achievement conducted by the Korea Institute of Curriculum and Evaluation, since 2003, the division score by achievement level has been established, and the trend of changes in academic achievement has been identified by linking the division score by achievement level. In educational achievement assessments, the cutoff points that differentiate achievement levels require evidence of the usefulness of test scores to support test score validity. Messick's (1989) definition of validity and Kane's (1992, 2006, 2013) argument-based approach to validation focus on test results and their use. If large-scale data collection is practically difficult, an alternative may be to link it with the Common European Framework of Reference for Language (CEFR), an internationally used external scale, and to compare the meaning of scores between tests indirectly.

As an existing scale for test scores that measure foreign language proficiency, the CEFR scale is widely used. CEFR helps people learn a foreign language independently according to their needs and helps language education institutions objectively evaluate by presenting consistent teaching contents and methods. Comparing and analyzing English tests developed overseas is possible, thereby increasing the utilization of scores. Suppose it is difficult to collect large-scale data for linking scores between different tests. In that case, an alternative is to set levels by linking them to an internationally used external scale such as CEFR, and indirectly compare the meaning of scores between tests.

Most of the English proficiency assessments researched CEFR relevance, and the Educational Testing Service (ETS) announced the relevance results of TOEFL iBT and TOEIC with the Common European Language Standards in 2008 (Tannenbaum & Wylie, 2008). To align with CEFR in the areas of speaking, writing, listening, and reading of TOEFL iBT and TOEIC, a manual that follows the establishment of relevance with the Common European Framework of Reference for Languages (CEFR) was used. Several studies on the relationship between the International English Language Testing System (IELTS) and CEFR have been conducted, and the results of a recent level-setting study were published in 2013 (Lim et al., 2013).

In South Korea, many studies have been conducted to prove the validity of the internationally verified CEFR by comparing it with the achievement standards of the Korean curriculum. The CEFR is a recognized language proficiency measure and was compared and analyzed with achievement standards for the first year of high school English. When developing achievement standards, it is argued that activities, examples, and evaluation examples of the standards should be presented together to make them actionable (Lee & Kim, 2009). Hwang (2016) argued that CEFR is a global standard for foreign language education and has high educational efficiency because it can diagnose and evaluate mermaid skills based on international common standards.

This study aims to link the Gtelph Level 2 exam with the CEFR. The split score was calculated using the improved Angoff method to match the common European language standard scale from A1 to C2, and the validity was reviewed by evaluating each method according to procedural, internal, and external criteria.

II. Theoretical background

1. Common European Framework of Reference for Languages (CEFR)

The Common European Framework of Reference for Languages (CEFR) is widely used to measure foreign language proficiency in education. In the 1990s, intuitive, qualitative, and quantitative methods were all used to develop the levels and skills for each level of the Common European Framework of Reference for Languages. CEFR provides a metalanguage and a reference point to various parties, such as foreign language professors, learners, evaluators, and textbook developers, to satisfy their efforts and the needs of learners, such as educational institution managers, textbook writers, teachers, teacher trainers, and test writers. It provides a way to introspect.

CEFR has six ability levels from A1 to C2 and presents detailed achievement level skills for each area of speaking, writing, listening, and reading and detailed skills in each area. In CEFR, six levels of A1, A2, B1, B2, C1, and C2 are standard, and achievement level descriptions for Pre-A1, A2+, B1+, and B2+ levels are also included. A1 and A2 indicate the basic level, and B1 and B2 mean the level at which the language can be used independently (independent). C1 and C2 represent proficient levels.

When CEFR is introduced, educational goals based on objectivity, transparency, and commonality of language education are set, and global and universal application standards are established in tests that measure foreign language proficiency so that learners' levels can be identified. In addition, CEFR includes dozens of scales in the Common European Framework of Reference for Languages, scales for detailed abilities under each domain such as speaking, writing, listening, and reading, various communication strategies, linguistics, sociolinguistics,

and pragmatic details. There are scales of ability, etc. It avoids education that judges language ability only with the overall score and presents tasks that learners can do in detail so that skill description questions are subdivided. The ability description questions are written in three categories: communication activities, strategies, and skills. It is described (Kim, 2019).

<Table 1> Descriptors of the Six Levels of CEFR

A1	The most basic level is where one can understand and use the most familiar and fundamental expressions in everyday life. For example, one can introduce themselves and ask about basic details of someone else, but communication is only possible if the other person speaks very slowly and helps.
A2	At this level, one can understand and use frequently used sentences and expressions in daily situations. For instance, they can exchange information about family, shopping, and nearby areas, and describe their background and surroundings with simple expressions.
B1	One can understand the main contents of familiar speech and writing encountered at work, school, and leisure activities. They can cope in places where the target language is spoken, briefly discuss their interests, and express experiences, events, dreams, and hopes, and can also give simple reasons for opinions and plans.
B2	This level denotes an understanding of abstract and complex types of speech and writing in one's field of expertise. They can easily communicate with native speakers and express opinions on various topics.
C1	At this level, one can understand various types of long and challenging speech and writing, including their implicit meanings. Those at this level are considered to be able to use the target language fluently and spontaneously for social, academic, and professional purposes. They can write clear, well-organized, and detailed texts on complex subjects.
C2	This represents the highest level where one can understand almost anything they hear or read. They can integrate and restructure speech and writing from various sources, speak and write spontaneously, very fluently, and precisely, even conveying finer nuances in more complex situations.

2. Level-setting

Level setting, which is a method of setting classification criteria, determines one or more division scores in a test. The test results are classified into two or more categories using the split score calculated through level setting (Cizek & Bunch, 2011). The split score is set as a standard setting, and a statement about the test taker's performance ability belonging to each section is prepared to classify the score scale into six parts, which are CEFR standards. The

standard setting includes stating the ability of the test takers in each area to know or perform, along with the setting of the split score, and the classified results are used for significant decision-making.

The CEFR development process collects data for scaling systematically and balances and scales the most excellent skill description items from the process quantitatively. A quantitative verification procedure such as the Rasch model is attempted to secure the validity of how much the subjectivity of the participants was involved. Scaled scores for cut scores are based on averages and represent cross-panel summary statistics (mean, median, standard deviation, minimum, maximum, and standard error of judgment).

Berk (1986) compared level-setting methods such as the Ebel method, the Nedelsky method, and the Angoff method and evaluated the Angoff method as the method that achieved the best balance between suitability and practicality. The Angoff method has been modified into various forms, such as the modified Angoff method, the confirmed Angoff method, and the Yes/No method, and the modified Angoff method is the most widely used. The modified Angoff method has the panel repeat probability predictions several times, provides feedback in the middle of the round, etc., and draws consensus among the panels. The main features of the modified Angoff method are: 1) panel members have a standard definition of minimum ability holders, 2) panel members discuss each other's judgments, and 3) panel members are based on past trial results. It can be regarded as providing standard information to people.

3. Validity

There are three methods for assigning competency-descriptive items to different levels: intuitive, qualitative, and quantitative. The best approach to developing language scales

combines all three approaches (Council of Europe, 2001). The qualitative research method uses the split point method that classifies the G-Telp Level 2 test into six ability levels and calculates the split score using the Angoff method. The split score is determined according to the judgment of the expert panel members, and additional opinions are collected through rounds considering the standard deviation to determine the final split score. The quantitative research method reviews qualitative research results according to internal, external, and procedural criteria and checks whether panel members agree on the final split score.

Cizek and Bunch (2011, p.81) showed the evaluation factors of level setting by classifying them into intrinsic, extrinsic, and procedural aspects. The intrinsic consistency evaluates the validity of the intrinsic level setting, the intrinsic consistency of each panel member, the consistency of judgment among panel members, the consistency of the test-taker classification through the final split score, and the degree of consistent classification of other items types, content areas, and test-takers. Classification agreement by round and the size of the standard deviation were analyzed.

The validity of extrinsic leveling can be demonstrated by comparing the results of applying split scores or other criteria. Examine whether the final split score finally decided by the panel members is realistically appropriate, and compare the split score obtained through the conversion table based on other tests with the final split score calculated through level setting. Examining the distribution of test takers by level based on the final split score is also a method of verifying the validity of the level-setting result from an external aspect.

The validity of the procedural level setting is whether the purpose and process are clear, whether the procedure and data analysis are easy, whether the panel selection and training, whether the level setting procedure is reasonable and systematic, and whether the panel has confidence in the level setting process and final division score. It is determined by evaluating

whether or not it has. To this end, the entire process is described in detail and evaluated, from panel selection, data provision, training process, and panel evaluation of the process and results.

III. research method

In scheduling and planning CEFR research, the research director needs expertise in education and measurement, evaluation, standard setting, and scaling. We use the modified Angoff method that classifies into six ability levels to carry out the qualitative research method of level setting. The modified Angoff method draws consensus among panels by having the panel repeat probability predictions several times and providing feedback in the middle of the round. Panel members have a standard definition of minimum ability holders and discuss each other's judgments to calculate the split score. The split score is determined according to the judgment of the expert panel members, additional opinions are collected, and a workshop is held to determine the final split score. The workshop consists of more than ten people, including professors, learners, evaluators, textbook developers, educational institution managers, textbook authors, and lecturers, and discussions are held to exchange opinions. Members can mainly include people who speak English as their mother tongue and experts on the test (e.g., Koreans). In this study, native speakers of Gtelp Korea and external researchers participated as a panel.

The study was conducted based on the CEFR manual, "Relating Language Examinations to the Common Framework of Reference for Languages : Learning, Teaching, Assessment (CEFR): A manual (2009)" Cumbersome procedures were excluded, and related documents are included in the manual. Utilized. A pre-orientation was held so that the expert

panelists could be fully aware of the entire process by sufficiently explaining the preparation, progress, and explanation of level setting.

1. Inspection tool

In this study, 4,627 participants in 2022 analyzed the G-telp Level 2 test results. Test takers included ordinary people, university students, and middle and high school students, and the test consisted of listening comprehension, grammar, and reading comprehension. The test consisted of 4 multiple-choice questions with a total of 80 questions, and the reliability of the test items was confirmed through a reliability test and a basic statistical test. <Table 2> shows the basic statistics for the test raw scores.

<Table 2> Descriptive Statistics

Domain	Number of Questions	Number of Test Takers	Average	Standard Deviation	Skewness	Kurtosis	Minimum Value	Maximum Value
G-telp Level2	80	4627	46.3	9.87	0.23	-0.82	15	91

2. Level-setting procedure

To set the level, experts in language evaluation, English education, applied linguistics, and experts with experience in English education and evaluation were gathered. A research representative and panel were formed. The panel size was determined to be at least ten according to the standard of Tannenbaum and Cho (2014), and efforts were made to produce

stable level-setting results by providing appropriate and diverse perspectives. Panel members were classified according to gender, nationality, native language, level setting experience, English teaching experience, and English evaluation experience.

A. Composition of the Expert Panel

The panel comprised 10 people, including learners, measurement experts, textbook writers, educational institution workers, textbook authors, and instructors. The participants were four native speakers from the internal global research team, two internal researchers, one internal textbook researcher, one native speaker professor, one native speaker instructor, and one internal and external native speaker researcher. Panel members were sent two weeks before the actual setting of levels, including data on common standards for European languages, setting levels, and schedules. The materials sent are as follows.

- Common standards for European languages, European languages (Council of Europe, 2001)

- Sister edition of Common European Language Standards (Council of Europe, 2018)

- European language common standards linkage manual (Council of Europe, 2009)

- Collection of Common European Language Scales related to grammar, reading comprehension, and listening comprehension

- A collection of grammar, reading comprehension, and listening comprehension questions linked to the Common European Language Standards Scale

- Linking ETS to Common European Language Standards for TOEFL and TOEIC

Research (Tannenbaum & Wylie, 2008)

- Background variable questionnaire

- Survey questionnaire

- CEFR vocabulary for listening comprehension and reading comprehension questions

- Level-setting schedule

Before participating in the level setting, panel members reviewed the materials sent in advance to familiarize themselves with the CEFR and level setting. CEFR presents subdivided scales for each subdomain of language ability, and its detailed ability and only scales related to reading comprehension were collected and provided to help panel members understand. In addition, other experts sent a collection of questions organized by level of reading comprehension questions developed or linked to the Common European Language Scale so that panel members could compare their understanding of the scale and make their judgments.

Before participating in the level setting, panel members reviewed the materials sent in advance to familiarize themselves with the CEFR and level setting. CEFR presents subdivided scales for each subdomain of language ability, and its detailed ability and only scales related to reading comprehension were collected and provided to help panel members understand. In addition, other experts sent a collection of questions organized by level of reading comprehension questions developed or linked to the Common European Language Scale so that panel members could compare their understanding of the scale and make their judgments.

The level setting was done over three weeks and three days from the third week of December 2022. On the morning of the first day, panel members gathered and were informed about the study's purpose and introduced the test's purpose, details of the test, and example questions. In addition, for the modified Angoff method, a level-setting method used in this study, the panel members were divided into three groups considering gender and background.

Each group conducted group activities to understand each area's Common European Language Standards. The panel members were given the Common European Framework of Reference for Languages and discussed in groups to adapt to the level-setting method and understand the concepts. In each area, the minimum ability holder was defined for each level of CEFR. After group activities, the definition of the minimum ability holder presented in each group was shared by all panel members and discussed.

In this study, because the number of test items in each domain is relatively small, the split score calculated using the modified Angoff method was calculated. Panel members were able to use more diverse feedback data to review the test from multiple angles and activate group discussions. In addition, if the panel judged that the test was unsuitable for test takers belonging to a certain level of the Common European Language Standards, it was guided to indicate "not applicable (N/A)" instead of presenting a split score. Finally, when more than 1/3 of the panel judged N/A, it was notified not to calculate the split score for that level.

<Table 3> Characteristics of Level Setting Panel

Category		Frequency	Percentage(%)
Gender	Male	6	60%
	Female	4	40%
Nationality	South Korea	3	30%
	USA	5	50%
	UK	1	10%
	Pakistan, USA	1	10%
Mother Tongue	Korean	3	30%
	English	5	50%
	English, Chinese	1	10%
	Pakistani	1	10%

Experience in Level Setting	Yes	2	20%	
	No	8	80%	
Experience (Years)	**Average**	**Standard Deviation**	**Minimum**	**Maximum**
English Education	7.2	5.3	0	15
English Assessment	4.5	3.5	1	10

A. How to proceed

The level setting was carried out for a total of 3 days. After the panel members gathered on the morning of the first day, information on the purpose of the study, the purpose of the test, details of the test, and example questions were introduced. The schedule was from 9:00 am to 5:00 pm, and rounds 1-4 were held in each area.

- Day 1: Grammar

- Day 2: Listening

- Day 3: Reading

In order to facilitate discussion among the panel members, considering the gender and background of the members, the group was divided into one in-house group and one external online group, and each group conducted group activities to understand the common European language standards.

Through group discussion, the minimum ability for each level of CEFR was defined, and then the minimum ability presented by each group was shared by all panel members to embody the concept and adapt to the level-setting method. Panel members reviewed the test from various angles using various feedback materials and tried to activate group discussion. To

apply the modified Angoff method, the corresponding round test sheet was printed as it is, and the correct answer was marked and provided. Six division points need to be determined to distinguish the six levels of CEFR. Still, it is burdensome for panel members to present six division scores in each round, so referring to the method of Tannenbaum and Wylie (2008), The judgment on the C2 level was made in rounds 1 and 2, and the judgment on the levels of A1, B1, and C1 was made in rounds 3 and 4.

In the first round, panel members were given sufficient time to review the test papers. In the case of the modified Angoff method, panel members judged the probability that the minimum ability holder of each level was correct for all questions from the first round. Moreover, from the second round, it was configured to evaluate each item with the sum of the probability of getting it right, that is, the total score. In each round, panel members were allowed to make independent judgments without discussion, and after the round was over, feedback data was calculated, and panel members took a break. Guides were provided on the content and meaning of each material, and after sufficient group discussion based on this, the next round began. The test paper used in the study provided feedback data by analyzing the performance of actual test takers. In addition, a unique number was assigned to each panel member, and a unique number was used instead of real names when providing feedback so as not to feel pressured to revise the judgment on the division score during group discussion.

In this study, we aimed to use the split scores derived from the modified Angoff method as the final result, and to ensure validity, we compared these results with those from the Bookmark method. By applying both methods simultaneously, panel members were encouraged to review the test from different perspectives, using a variety of feedback materials, which helped foster active group discussions. However, to avoid influencing the results, we avoided disclosing the research objectives prematurely to the panel. Both methods for

determining the split scores were described equally without indicating which would be used for the final decision.

For the modified Angoff method, examinees' test sheets from the respective round were provided as-is, with correct answers marked. In the Bookmark method, a sequence set of 80 items was created from two test sheets conducted around the same time. This was done because having fewer questions in a sequence tends to make the scale intervals for each question appear broader. Generally, a test sheet with an item sequence set has fewer very easy or very difficult questions and more questions of medium difficulty, resulting in larger difficulty gaps at both extremes. To minimize judgment errors by the panel members, we followed Cizek and Bunch's (2011, p. 228) suggestion to add questions with very low or very high difficulty from other test sheets to reduce these gaps.

A two-parameter logistic model was used to estimate the item parameters, and the questions in the item sequence set were organized based on both difficulty and discrimination parameters. Additionally, for each question in the set, the original question number from the test sheet, the ability level at which a 2/3 probability of answering correctly was achieved, and the correct answer were provided. For the 1-passage-2-question format, an indication of which questions shared the same passage was also included at the bottom of the sequence set.

A. Details of the feedback provided to the panel after each round.

1) After the 1st round (Common European Language Standards A2, B2, C2)

- Round 1 split score presented by each panel member by the two-level setting method

- Average, median, and standard deviation of round 1 split scores of all panel members

- P-values of items for the entire group, top 25%, and bottom 25%

2) After Round 2 (Common et al. A2, B2, C2)

- Round 2 split score presented by each panel member by the two-level setting method

- Average, median, and standard deviation of round 2 split scores of all panel members

- Percentage of test takers divided by the average of the second round split points

- Result of linking European Language Common Criteria indirectly calculated through conversion tables with other tests

3) After three rounds (Common European Language Standards A1, B1, C1)

- The 3rd round split score presented by each panel member by the two-level setting method

- Average, median, and standard deviation of the third-round split scores of all panel members

- P-values of items for the entire group, top 25%, and bottom 25%

1) After four rounds (Common European Standards A1, B1, C1)

- The 4th round split score presented by each panel member by the two-level setting method

- Average, median, and standard deviation of the fourth-round split scores of all panel members

- Percentage of test takers divided by the average of the third and fourth round split points

- Result of linking European Language Common Criteria indirectly calculated through conversion tables with other tests

3. Validation method

Calculate Cohen's Kappa coefficient and classification accuracy coefficient, respectively, to analyze the classification agreement and classification accuracy according to the final split score as an internal criterion for verifying the standard deviation of the split scores presented by each panel member and verifying the validity of level setting (Hanson & Brennan, 1990; Lee et al., 2002). The classification accuracy coefficient is a coefficient that indicates how accurate the classification through the observation score is when considering the accurate score and the probability that an examinee whose actual true score τ is lower than the true split score $\tau 0$ in the whole will get a score higher than the observed split score x0 (false positive rate) and the probability that an examinee whose actual true score τ is higher than the true split score $\tau 0$ will obtain a score lower than the observed split score x0 (false negative rate) (Hanson & Brennan, 1990). This study calculated classification agreement and accuracy based on item response theory (Hanson & Brennan, 1990; Lee et al., 2002). In other words, the actual score distribution and error score distribution are obtained using the item parameter and ability parameter distributions estimated from actual data, and based on this, the Kappa coefficient and classification accuracy coefficient are calculated for the split score calculated in this study (Lee & Kim, 2009) was obtained. The Jamovi 2.3.2 (Seol, 2022) program was used to calculate the coefficients in this study. As a criterion for external validity, the distribution of test takers by the level of the Common European Language Standard was calculated according to the result of the final split score. Finally, to secure procedural standards for evaluating the validity, the calculated split score was reviewed after completing up to 4 rounds, and the level-setting process was evaluated by responding to questionnaires. Questions included asking about the

evaluation of the level-setting process, the factors that influenced the judgment, and the confidence in the final split score.

IV. results

The level setting result of the G-telp level 2 test is the same as <Table 4>. The split scores calculated using the modified Angoff method are the modified Angoff 1st and second round A2, B2, and C2 split scores and the third and fourth round A1, B1, and C1 split scores in bold.

In the final split score calculation process, all 10-panel members presented a split score other than N/A for all Common European Language Levels. Also, there were no outliers, so the average was used without considering other statistics. To verify the internal validity, the standard deviation presented by each panel member was investigated.

<Table 4> shows the results of level setting according to the modified Angoff method. During rounds 1-4, the standard deviation between the split scores judged by each panel decreased and showed a tendency to converge in a specific direction. At most levels, we found that the average difference decreased as the rounds progressed. The split score for A2 and B2 decreased, and the split score for B2+ tended to increase.

"To assess the validity of the level setting from an external standpoint, a comparison between the results from the modified Angoff method and the Bookmark method shows that the average difference in split scores between the two methods is not significant for each round. Furthermore, as the rounds progress, the average difference tends to diminish across most levels."

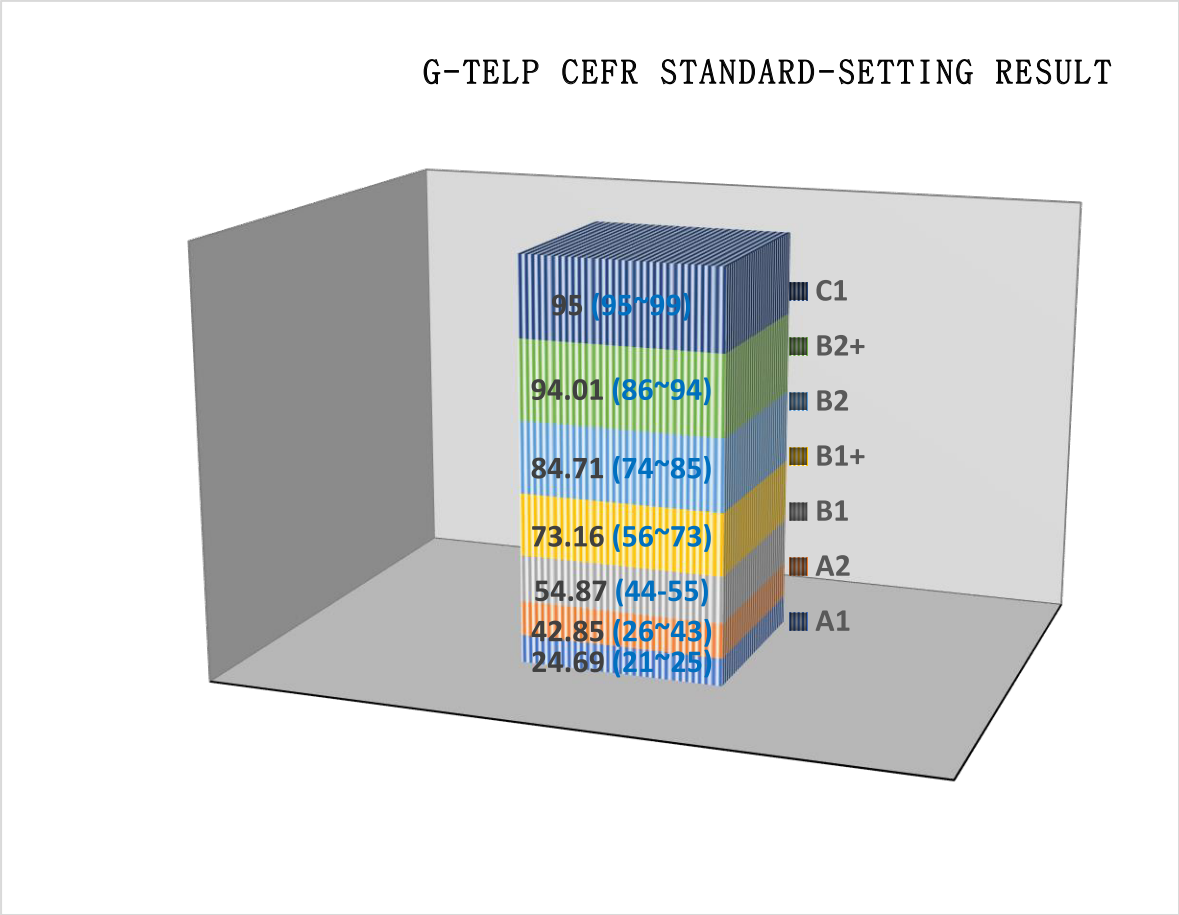
<Table 4> Results of Level Setting Based on the Modified Angoff Method

	CEFR	Round 1 Average(SD)	Round2 Average(SD)	Round 3 Average(SD)	Round 4 Average(SD)
Modified Angoff	A1	-	-	24.91(4.69)	**24.69(3.60)**
	A2	43.36(6.85)	**42.85(5.33)**	-	-
	B1	-	-	56.53(5.98)	**54.87(4.96)**
	B1+	-	-	74.88(10.12)	**73.16(9.97)**
	B2	87.50(9.81)	84.71(7.87)	-	-
	B2+	93.54(5.27)	94.01(4.10)	-	-
	C1	-	-	**95~**	-
Bookmark	A1	-	-	25.98(4.37)	**25.77(53.45)**
	A2	44.16(6.75)	**43.65(4.93)**	-	-
	B1	-	-	57.13(5.38)	**55.16(4.91)**
	B1+	-	-	75.13(9.12)	**73.79(9.36)**
	B2	88.45(9.31)	85.25(7.15)	-	-
	B2+	94.74(5.78)	95.51(3.71)	-	-
	C1	-	-	**95~**	-

Table 5> summarizes the result of the level setting by the modified Angoff method in <Table 4> as a CEFR grade comparison table as the result of the grade comparison table according to the European language common standards according to the final round.

<Table 5> G-TELP CEFR Level Comparison Table: Results of Level Setting Based on the Modified Angoff Method

CEFR	G-TELP 2 Level	2, 4 라운드 평균	표준편차
A1	**21-25**	24.69	5.60
A2	**26-43**	42.85	5.33
B1	**44-55(B1)** **56-73(B1+)**	54.87 73.16	6.96 9.97
B2	**74-85(B2)** **86-94(B2+)**	84.71 94.01	7.87 4.10
C1	**95~99**	-	-
C2	**-**	-	-



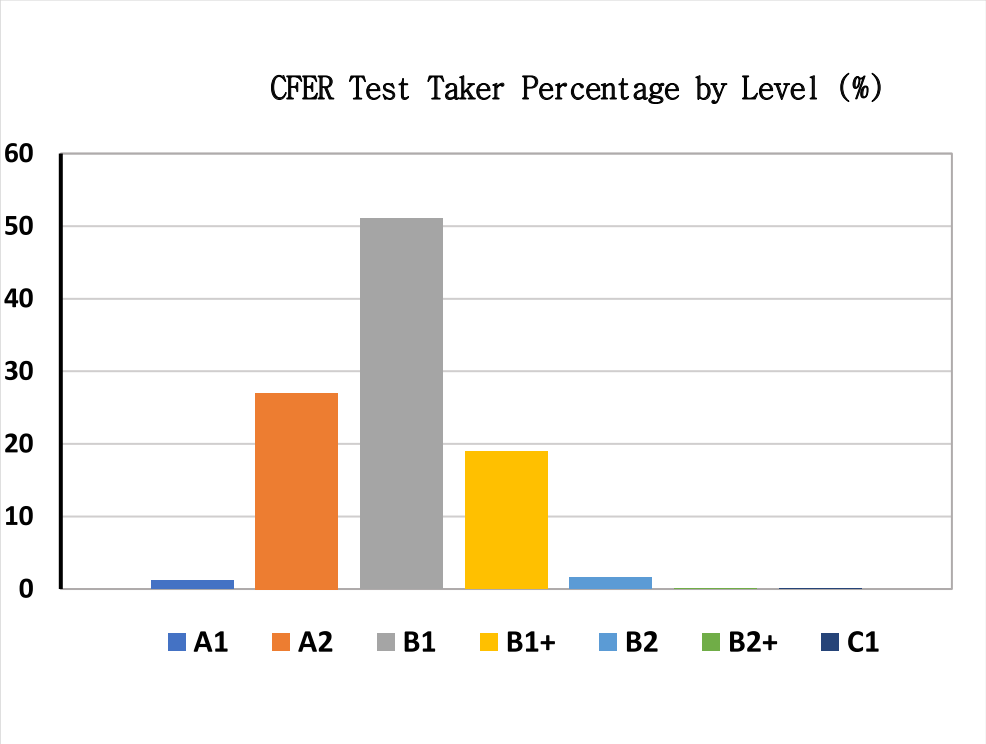
<Figure 1> G-telp CEFR Grade Table: Level Setting Results According to the Modified Angoff Method

<Table 6> shows the distribution of test takers by level according to the result of the modified Angoff level setting, which was presented as a criterion for external validity of the level setting result. It shows the distribution by the level of the test takers in the round, which is based on the result of the modified Angff level setting and external validity. As a result of the analysis, it was found that about 51% of the examinees who participated in the trial round met the Common European Language Standard B1. A2 and B1+ level test takers accounted for 27% and 19%, respectively, and B2+ level test takers accounted for only about 0.10%. C1 was 0.02%, and no test takers with C2 level were found. In the recent G-Telp Level 2 test trend, some test takers scored 95 points or higher or got a perfect score, so they are judged to be at the C1 level.

<Table 6> Distribution of test takers based on the CEFR levels according to the modified Angoff standard-setting results

CEFR	Proportion of test takers(%)
A1	1.20
A2	27.00
B1	51.00
B1+	19.00
B2	1.60
B2+	0.10
C1	0.02
C2	0.00

* In total, 99.82%, with the remaining 0.18% being the percentage of test takers below the A1 level



<Figure 2> Distribution of test takers based on the CEFR levels according to the modified Angoff standard-setting results

<Table 7> summarizes the questions related to level setting. As a result of the analysis, 99% of the survey subjects evaluated that the preliminary data provided before participation in the level setting was valuable, and 100% responded that they understood the purpose of the study. In addition, all survey subjects evaluated that the facilitator's instructions and explanations were clear, the explanation of the level-setting method was detailed, and the division score calculation method was straightforward. The feedback information and discussions provided after each round of level-setting were evaluated as useful, and the process of making level-setting judgments was easy.

<Table 7> Survey Response Results: Level Setting Process

Q. How strongly do you agree or disagree with the following statements?				
	Strongly agree	**Agree**	**Disagree**	**Strongly Disagree**
The homework assignment was useful preparation for the study.	25%	63%		
I understood the purpose of the study.	75%	25%		
The instructions and explanations provided by the facilitators were clear.	58%	42%		
The training in the standard-setting methods was adequate to give me the information I needed to complete my assignment.	29%	58%		
The explanation of how the recommended cut scores were computed was clear.	25%	75%		
The opportunity for feedback and discussion between rounds was helpful.	63%	38%		
The process of making the standard-setting judgments was easy to follow.	54%	46%		

* Due to missing data and multiple responses, the total is less than or exceeds 100%.

<Table 8> summarizes the factors influencing the level of judgment among the survey questions. The most influential factors were my professional experience (60%), my definition of minimum ability (52%), and group discussions between rounds (52%). The response that they were influenced by the split score presented by other panel members was relatively low (32%).

<Table8> Survey Response Results: Factors that Influenced Level-setting Judgments

Q. How influential was each of the following information sources on your cut-score decisions?			
	Very influential	**Somewhat influential**	**Not influential**
The definition of the minimally competent person	52%	32%	16%
The between-round discussions*	52%	24%	12%
The cutscores of other panel members	32%	56%	0%
My own professional experience*	60%	40%	0%

<Table 9> shows the survey results on the confidence or certainty of the panel members on the final score of each level of the Common European Language Standards. The panel who participated in the setting of the G-telp level 2 showed high confidence in A1 and C1 but relatively low confidence in the A2 and B2 division scores.

<Table 9> Survey Response Results - Confidence in the Final Split Score

Q. How comfortable are you with the final cut score recommendations established by the panel? (Circle one)				
	Very comfortable	**Somewhat comfortable**	**Somewhat uncomfortable**	**Very uncomfortable**
Cut score for CEFR A1	29.17%	66.67%	4.17%	0.00%
Cutscore for CEFR A2	16.67%	70.83%	8.33%	0.00%
Cut score for CEFR B1	20.83%	75.00%	4.17%	0.00%
Cut score for CEFR B2	16.67%	70.83%	8.33%	0.00%
Cut score for CEFR C1	29.17%	62.50%	8.33%	0.00%

Table 10> shows the classification consistency and accuracy for the split scores derived after each round as an internal criterion for evaluating the validity of the level setting. At this time, the Kappa coefficient was used to indicate the degree of classification agreement. As a result of the analysis, it was found that the classification concordance and classification accuracy for the three division scores (A2, B2, C2) gradually increased as the process progressed from round 1 to round 2. In the modified Angoff method, classification concordance and classification accuracy continuously increased in rounds 1 and 2, but classification accuracy and classification concordance decreased slightly in round 3. However, classification agreement and accuracy increased in the third and fourth rounds. In the third round, the case of classifying six levels with a total of six division points was analyzed by adding the division scores for A1, B1, and C1 to the three division scores calculated up to the second round. As the number of levels increases, the classification agreement and classification accuracy decrease. Hence, the results were lower than those of the 1st and second rounds. However, considering the range observed in previous studies and the number of levels to be classified in this study,

the consistency and classification accuracy were high.

<Table 10> Classification Agreement and Classification Accuracy Coefficients for Split Scores in Each Round

		Round 1	Round 2	Round 3	Round 4
Modified Angoff	Classification Agreement	0.554	0.581	0.532	0.543
	Classification Accuracy	0.812	0.830	0.792	0.802
Bookmark	Classification Agreement	0.542	0.554	0.510	0.514
	Classification Accuracy	0.798	0.812	0.742	0.752

V. Conclusions

In this study, an expert panel was formed to link the US ITSC's G-telp English Proficiency Test with the CEFR scale, and the split score was calculated using the modified Angoff method. As a result, it was possible to obtain division scores for all areas of G-telp that divide the six CEFR levels. Dot product to support the validity of this leveling study. Cross product. Procedural standards have been secured in the following aspects. Internally, the standard deviation of the split scores judged by the expert panel members tended to decrease as rounds passed. The classification agreement and classification accuracy coefficient for the split score also showed good values, verifying the validity of the split score. Externally, according to the result of setting the level, the distribution of test takers by the level of the expected standard for European languages appeared appropriate. Procedurally, the panel's characteristics and the entire level-setting process were described. Most of the panel members

evaluated the level-setting preparation and progress, and the explanation was clear and helpful when making a judgment. In addition, most panel members responded that they were confident in the final split score.

The final split scores obtained from both level-setting methods were similar; however, at each level, the split scores generated using the Bookmark method were higher than those from the modified Angoff method. This finding is consistent with results from earlier studies (Lee Kyumin et al., 2014; Buckendahl et al., 2002). Additionally, in terms of classification consistency and accuracy, the modified Angoff method produced more consistent results than the Bookmark method, which aligns with previous research findings (Kim Sun et al., 2009).

As a result of the study, it was found that the G-telp Level 2 test can be used to discriminate the six competency levels of the CEFR scale. Panel members judged they could classify from the most basic A1 level to the highest C2 level on the CEFR basic scale. The test can measure various abilities, including items of varying difficulty. The grade comparison table based on the CEFR scale obtained through analyzing the results after the workshop appears to be comparable with other English proficiency tests. This served as an opportunity to verify that it was not significantly different from the past G-telp CEFR class comparison table. In addition, the results of the company's CEFR workshop by deriving grades of B1+ and B2+ are of great significance as they show a more advanced form than the previous results.

In this study, to evaluate the contents and methods of G-Telp consistently and objectively to support its validity, we conducted a linkage with CEFR, the latest standard for language education, and investigated the application method. The introduction of the CEFR indicates that the emphasis on sociolinguistic knowledge and language strategies, rather than a teaching method focusing on grammar and vocabulary, was considered in the G-telp Level 2 exam. Referring to the level setting schedule and progress established in this study, it is thought

that other G-telp tests (Speaking et al. 3, G-telp Junior, Etc.) can be linked with CEFR. Based on the results of this study, it is possible to study the level setting of the Common European Language Criteria by comparing the level of difficulty in different questions of the same test, along with setting the level in connection with CEFR and increasing the number of panel participants more diversely. In addition, it is necessary to conduct a study on the equivalence of other English-speaking tests and the G-telp speaking test and to find a way to understand the meaning of scores by applying this as the setting of the level of linkage to the Common European Language Standards and to look into follow-up studies.

VI. References

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*, pp. *56*, 137–172.

Cizek, G. J., & Bunch, M. B. (2011). *준거설정 (성태제 역)*. 서울: 학지사. (원서출판 2007).

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*. Strasbourg, France: Council of Europe.

Council of Europe. (2018). Common European Framework of Reference for Languages: Learning, teaching, assessment. *Companion volume with new descriptors*. Strasbourg, France: Council of Europe.

Hanson, B. A., & Brennan, R.L.,(1990). An investigation of classification consistency indexes estimated under alternative strong accurate score models. *Journal of Educational Measurement*, 27, 345 – 359.

Hwang, P.-A. (2016). Adequacy of primary English reading achievement standards in the 2015 revised national curriculum. The Korea English Education Society, 15(4),147-165.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527-535.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 17-64). Westport, CT: Greenwood Publishing

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1-73.

Kim, H. S. (2019). *Foreign language teaching in Korea concerning CEFR. Institute for Humanities and Social Sciences*, 20(4), 79-96.

Lee, Y. S., & Kim, H. Y. (2009). *A comparative study of the achivement standards between the revised Korean national curriculum of English and Common European Framework of References (CEFR).* Modern English Education, 10(2), 108-132.

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*, 32-49.

Messick, S. J. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York:Macmillan.

R Core Team (2021). *R: A Language and environment for statistical computing*. (Version 4.1) [Computer software]. Retrieved from https://cran.r-project.org. (R packages retrieved from MRAN snapshot 2022-01-01).

Seol, H. (2022). *snowRMM: Rasch Mixture Model for jamovi*. [jamovi module]. Retrieved from https://github.com/hyunsooseol/snowRMM.

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 Research Report 08-34). Princeton, NJ: Educational Testing Service.

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, 233–249.

The jamovi project (2022). *jamovi*. (Version 2.3) [Computer Software]. Retrieved from https://www.jamovi.org.

Willse, J. (2014). *mixRasch: Mixture Rasch Models with JMLE*. [R package]. Retrieved from https://CRAN.R-project.org/package=mixRasch.

Appendix A: <Round table schedule>

AGENDA: Mapping G-TELP Test Onto the

Common European Framework

Day 1: G-TELP Grammar Section

Day 2: G-TELP Reading Section

Day 3: G-TELP Listening Section

Day 4: G-TELP Speaking/Writing Section

9:00 – 10:20 Table Groups: groups for A2, B2, and C2 for Grammar/Reading/Listening

Review of A1, B1, and C1 descriptions,

Overview of selected-response standard-setting methods/practice

10:30 – 10:40 Break

Angoff- method

10:40 – 11:20 Individual Round 1 judgments (A2, B2, C2)

11:20 – 11:30 Break (data entry)

11:20 – 12:00 Discussion of Round 1 results and score information

12:00 – 13:00 Lunch

13:00 – 13:30 Round 2 individual judgments (A2, B2, C2)

13:30 – 13:40 Break (data entry)

13:40 – 14: 00 Discussion of Round 2 results and impact data

14:00 – 14:20 Round 3 individual judgments (A2, B2, C2)

14:20 – 14:30 Break (data entry)

14:30 – 15:00 Round 4 judgments of A1, B1, and C1 relative to judgments for A2, B2, and C2

15:00 – 15:10 Break (data entry)

15:10 – 15:30 Discussion of Round 4 results and score information

15:30 – 15:50 Final individual judgments for A1, B1, and C1

15:50 16:00 Break (data entry)

Bookmark method

16:00 – 16:20 Individual Round 1 judgments (A2, B2, C2)

16:20 – 16:30 Discussion of Round 1 results and score information

16:30 – 16:50 Round 2 individual judgments (A2, B2, C2)

16:50 – 17:00 Discussion of Round 2 results and impact data

17:00 – 17:20 Round 3 individual judgments (A2, B2, C2)

17:20 – 17:40 Discussion of A1, B1, and C1 relative to judgments for A2, B2, and C2

17:40 – 17:50 Final individual judgments for A1, B1, and C1

Appendix B: Background Questionnaire

Name : _____ Participant # _____

1. What is your gender?

Male

Female

Other (please elaborate if you feel comfortable doing so)

_____

2. What is your nationality?

Province/state: _____

Country: _____

3. What is your first language (mother tongue)?

Language: _____

4. Do you speak any other languages?

Yes

No

If yes, please list each language and your approximate level (beginner, intermediate, advanced). Language 1: _____ Level:

_____

Language 2: _____ Level:

_____

5. Do you have any experience for Standard-setting?

_____

6. How many years do you have English education experiences (year)?

_____

7. How many years do you have English evaluation/measurement experiences? (year)

_____

Appendix C: Survey Questionnaire

Date:

Session: Grammar/ Listening / Reading / Speaking / Writing

1. Standard-setting process

How strongly do you agree or disagree with each of the following statements?

	Strongly Agree	**Agree**	**Disagree**	**Strongly Disagree**
The homework assignment was useful preparation for the study.				
I understood the purpose of the study.				
The instructions and explanations provided by the facilitators were clear.				
The training in the standard-setting methods was adequate to give me the information I needed to complete my assignment.				
The explanation of how the recommended cut scores were computed was clear.				
The opportunity for feedback and discussion between rounds was helpful.				
The process of making the standard-setting judgments was easy to follow.				

2. Factors influencing level setting judgment

How influential was each the following information source on your cutscore decisions?

	Very Influential	Somewhat Influential	Not Influential
The definition of the minimally competent person			
The between-round discussions			
The cutscores of other panel members			
My own professional experience			

Others:

_____

_____

3. Confidence in final cutscore

How comfortable are you with the final cutscore recommendations established by the panel?

	Very Comfortable	Somewhat Comfortable	Somewhat Uncomfortable	Very Uncomfortable
Cutscore for CEFR A2				
Cutscore for CEFR B1				
Cutscore for CEFR B2				
Cutscore for CEFR C1				
Cutscore for CEFR C2				

4. Do you have any concerns about the way the workshop was conducted?

_____

Appendix D: Penel for Standard-setting workshop

Internal Panel

Minjung Kim

Hajoon Yoo

Sunny Jeong

Candice Bayley

Corey Steiner

Rob Walsh

Toby Charles William

External Panel

Ali Raddaoui

Kymberly Talor

Mike Dong