

Linking an English Language Test (GTELP Junior) to the CEFR

Embracing Foundational Teaching in the Age of AI

2025

Prepared by

Minjung Kim, PhD

Conference

The National Council on Measurement in Education(NCME) Annual Meeting

Table of Contents

- 1. Background**
- 2. Purpose of the Study**
- 3. Methodology**
 - Structure of G-TELP Junior
 - CEFR Mapping Procedure
- 4. Results**
 - CEFR Score Mapping
- 5. Conclusion**

Background

What is CEFR?

- CEFR stands for the Common European Framework of Reference for Languages.
- It is an international standard for describing language proficiency.

Levels:

- A1, A2: Basic user
- B1, B2: Independent user
- C1, C2: Proficient user
- Additional Descriptions: Pre-A1, A2+, B1+, B2+

Key Features:

- Provides objectivity, transparency, and commonality in language education goals.
- Used in global exams: TOEFL, IELTS, DELE, DELF/DALF, Trinity, and G-TELP
- Avoids relying solely on overall scores; gives detailed task-based skills.
- Scales for every language domain: Listening, Reading, Speaking, Writing.

Noted: Detailed tasks & skills in communication, strategies, and more (Kim, 2019).

Background

The Common European Framework of Reference for Language (CEFR)

Level	Summary Description
A1	Understands and uses basic everyday expressions. Can introduce oneself and ask/answer simple personal questions if the interlocutor speaks slowly.
A2	Understands frequent, familiar expressions in daily life (e.g., shopping, directions). Can describe surroundings and exchange simple information.
B1	Understands main ideas of familiar topics at work/school. Can handle travel, talk about experiences, interests, and give simple opinions.
B2	Understands abstract or technical texts in their field. Can interact fluently with native speakers and discuss a wide range of topics.
C1	Understands complex and nuanced language, both spoken and written. Can express themselves fluently, flexibly, and effectively in academic or professional settings.
C2	Understands virtually everything heard or read. Can summarize and reformulate information, and express themselves very precisely and naturally, even in demanding contexts.

Background

Level-setting in Testing

- **Purpose:** Classify test results using division scores.

- **Methods:**

- Sets division scores (cut scores).
- Classifies score scale into CEFR's six standards.
- Describes the test taker's performance ability for each category.

- **Standard Setting Process:**

- Balances and scales skill description items.
- Uses quantitative verification like the Rasch model.
- Scaled scores for cut scores based on summary statistics.

- **Level-setting Methods:**

- Ebel, Nedelsky, Angoff methods.
- Angoff method is seen as balanced and practical.
- Modified Angoff is the most popular: Defines standard minimum ability, facilitates panel discussions on judgments, and provides standard information based on past trial results.

- **Reference:** Cizek & Bunch, 2011; Berk, 1986.

Background

Validity in Assigning Competency-Descriptive Items

- **Evaluation Factors of Level Setting** (Cizek & Bunch, 2011):
 - **Intrinsic:** Consistency within and among panel members and with test-taker classifications.
 - **Extrinsic:** Comparing results with other criteria. Verifying the appropriateness of cut scores using external data, like other tests.
 - **Procedural:** Clarity of purpose and process, ease of data analysis, panel training, systematic procedure, panel confidence in results.
- **Key Takeaway:** Validity is ensured through thorough processes, panelist evaluations, and external comparisons.

Background

Overview of G-TELP Junior

- G-TELP Junior is designed for young learners (elementary to middle school)
- Tests include Listening and Reading sections.
- Organized into 5 levels tailored by age and difficulty.
- CEFR alignment enables international interpretation of scores

Background

Test Structure by Level

- Level 1: Middle school 2–3 | Listening: 30 | Expresses ideas in short, logical sentences.
 - Level 2: Elementary 6–Middle 1 | Listening: 25 | Reading: 30 | Basic short conversation.
 - Level 3: Elementary 4–5 | Listening: 25 | Reading: 25 | Word transformation comprehension.
 - Level 4: Elementary 3 | Listening: 23 | Reading: 17 | Understands sentences via short word combinations.
 - Level 5: Elementary 1–2 | Listening: 27 | Reading: 13 | Recognizes objects but cannot express independently.
-

Purpose

Implementation & Relevance of CEFR in Testing

CEFR's Widely Recognized Role:

- Enables comparison and analysis of English tests developed globally.
- ETS linked TOEFL iBT and TOEIC to CEFR in 2008 (Tannenbaum & Wylie, 2008).
- Studies on the relationship between IELTS and CEFR, with recent findings in 2013 (Lim et al., 2013)

•CEFR in South Korea:

- Studies prove the validity of CEFR against Korean curriculum standards.
- Emphasis on actionable achievement standards (Lee & Kim, 2009).
- Hwang (2016): CEFR is a global standard with high educational efficiency.

This Study's Goal: Link the G-TELP Junior exam to CEFR.

Validate through the improved Angoff method and reviewing procedural, internal, and external criteria.

Methodology

CEFR Research & Level Setting

Panel Composition:

- Over 10 members: Professors, learners, evaluators, textbook developers, managers, authors, and lecturers.
- Native English speakers and test experts, including Koreans.
- Participation of native speakers from Gtelp Korea and external researchers.

Inspection Tool:

- Analyzed G-TELP Junior test results in 2024.
- Test takers: elementary, middle & high school students.
- Test components: Listening comprehension, reading comprehension.

Methodology

Level Setting Execution:

- Spanned over 4 days of workshops in 2024
- Day 1:** Introduction to the study's purpose, test details, and example questions.
- Activities to understand European Language Standards; defined minimum ability holder for each CEFR level.
- Options to indicate "N/A" if a test was unsuitable for specific CEFR level S.

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

Methodology

Research Approach & Methodology:

- **Objective:** Validate split scores derived from the modified Angoff method
- **Procedure:**
 - Modified Angoff: Examinees' test sheets printed with correct answers.
- **Reasoning:** To address large difficulty gaps, especially for tests with fewer easy or difficult questions.
- **Additional Info:** Angoff: Each item's original number, 2/3 probability ability parameter, and correct answer were provided. Shared passages for the 1-passage-2-question type were indicated.

Methodology

CEFR Level Setting Procedure

Procedure:

- Morning briefing: Study's purpose, test details, and example questions.
- Daily schedule: 9:00 am - 5:00 pm, 4 rounds/day.
- Panels: Divided by gender & background (in-house & external online groups).
- Activities: Discuss & define minimum ability for CEFR levels.
- Method: Modified Angoff - mark correct answers; 6 division points.
 - Rounds 1 & 2: A2, B2, C2 Level
 - Rounds 3 & 4: A1, B1, C1 Levels

Results

Methods Compared:

- Modified Angoff Method

1.No outliers detected; average used for calculations.

2.Across rounds 1-4, standard deviation between panelists' split scores reduced, indicating convergence.

3.Most levels saw a decreasing average difference with each round.

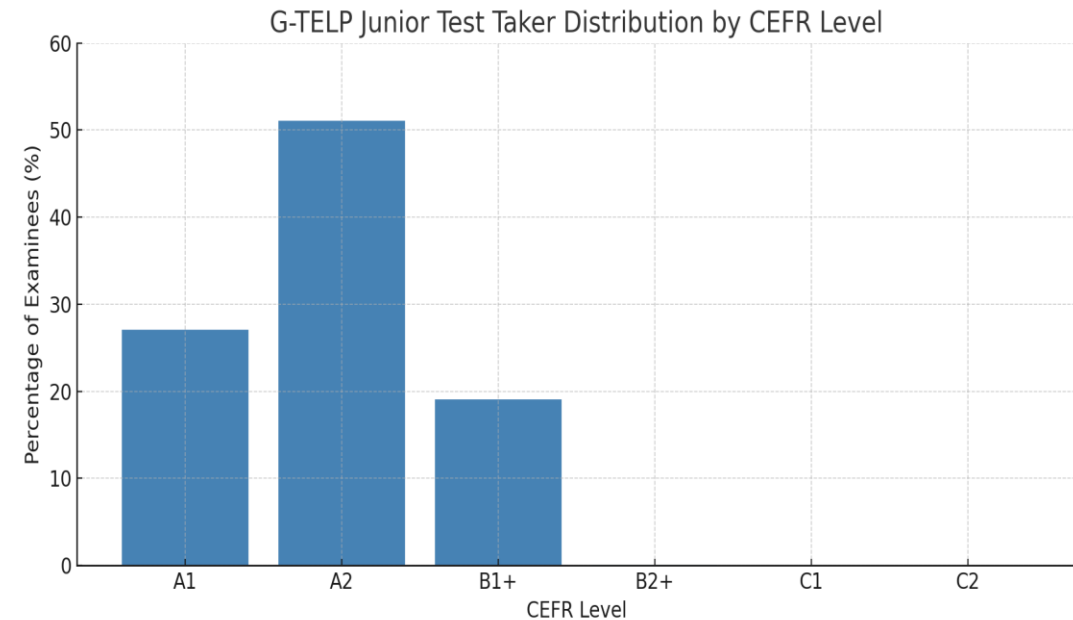
	Level 5	Level 4	Level 3	Level 2	Level 1
C2					
C1					
B2+					280.52
B2	292.11	289.64	285.39	277.22	273.52
B1+	280.91	275.63	267.02	270.12	255.26
B1	271.51	261.87	248.32	243.29	213.96
A2+	265.61	243.56	223.12	209.24	185.13
A2	245.26	212.99	185.85	172.44	140.74
A1	228.83	186.03	157.01	145.23	115.09

Results

- Assess external validity of level setting results.
- Understand the distribution based on the modified Angoff level setting.

Test Taker Distribution by Level:

- B1 Level:** ~51% of examinees
- A2 Level:** 27%
- B1+ Level:** 19%
- B2+ Level:** 0.10%
- C1 Level:** 0.02%
- C2 Level:** 0% (No test takers found at this level)



Results

Level Setting Process

1. Preliminary Data Evaluation: 99% found the provided preliminary data valuable.

2. Understanding of Study Purpose:

- 100% understood the purpose of the study.

3. Clarity of Instructions: All participants found the facilitator's instructions and explanations clear.

4. Feedback & Discussions:

- All participants found the feedback and discussions after each round useful.

<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.	50%	50%		
I understood the purpose of the study.	58%	42%		
The instructions and explanations provided by the facilitators were clear.	63%	38%		
The training in the standard-setting methods was adequate to give me the information I needed to complete my assignment.	75%	25%		
The explanation of how the recommended cut scores were computed was clear.	25%	75%		
The opportunity for feedback and discussion between rounds was helpful.	58%	42%		
The process of making the standard-setting judgments was easy to follow.	63%	38%		

Results

Factors that influential level-setting judgments

1. Professional Experience: 60% of participants indicated this as a major influence.

2. Personal Definition of Minimum Ability: Influenced 52% of the participants.

3. Group Discussions between Rounds: Equally influenced 52% of the participants

4. Split Scores from Other Panel Members: Relatively lower influence, with only 56% indicating it impacted their judgment.

<Survey Response Results: Factors that influenced level-setting judgments>

Q. How influential was each of the following information sources on your cutscore decisions?

	Very influential	Somewhat influential	Not influential
The definition of the minimally competent person	52%	32%	16%
The between-round discussions*	52%	32%	0%
The cutscores of other panel members	56%	24%	12%
My own professional experience*	60%	40%	0%

Results

Panel Confidence on Common European Language Standards:

•High Confidence Levels:

- **A1:** Panel showed strong confidence in this level's final score.
- **C1:** Similarly, a high degree of confidence was observed for the C1 level.

•Lower Confidence Levels:

- **A2:** Panel had relatively less confidence in the division score for A2.
- **B2:** Confidence in the B2 division score was also notably lower.

<Survey Response Results- Confidence in the Final Cut 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%
Cut score 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%

Results

Classification concordance and accuracy
for A2, B2, C2: Cohen's Kappa

- Round 1 to Round 2:** Gradual Increase
- Round 3:** Slight Decrease.
- Round 3 to Round 4:** Increase.

A slight dip was observed in Round 3, but recovery was noted in subsequent rounds.

As level numbers increased, both classification agreement and accuracy decreased, leading to slightly lower results in Round 3 compared to Rounds 1 and 2

References: The Jamovi project (2022).
(Version 2.3)

<Classification Agreement and Classification
Accuracy Coefficients for Cut Scores in Each Round>

		Round 1	Round 2	Round 3	Round 4
Modified Angoff	Classification Agreement	0.545	0.518	0.523	0.535
	Classification Accuracy	0.823	0.835	0.798	0.812

Conclusion

Linking G-TELP Junior to CEFR: Methods & Validity

- **Objective:** Link G-TELP Junior English Proficiency Test with CEFR.
- **Methodology:**
 - Formation of an expert panel.
 - Utilize the modified Angoff method for cut score calculation.
 - Ensured procedural standards for validity.

Conclusion

Implications, Comparisons & Future Research

- **Consistency with Previous Research:**

- Modified Angoff showed great classification consistency and accuracy

- **Results:**

- Obtained division scores for all G-TELP Junior areas across six CEFR levels.
- Decreased standard deviation of cut scores over subsequent rounds.
- High classification agreement and classification accuracy.

- **External Validation:**

- Appropriate distribution of test-taker levels for expected
 - European language standards.
-

Conclusion

Implications, Comparisons & Future Research

•Significance:

- G-TELP Junior can discriminate across all six CEFR competency levels.
- Test can measure a range of abilities with varying item difficulties.
- Comparison table with CEFR found comparable to other proficiency tests.
- Introduced advanced levels (B1+ & B2+).

•Future Directions:

- Link other G-TELP tests (Speaking et al. 3, G-TELP Junior, etc.) with CEFR.
- Integrating CEFR in G-TELP
- Explore the level setting of CEFR by comparing question difficulty.
- Investigate the equivalence of other English tests with G-TELP and study the meaning of scores linked to CEFR.

Thank you