

ITSC G-TELP

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**

Date **April 26, 2025**

Table of Contents

01

Background

02

Purpose of the Study

03

Methodology

Structure of G-TELP Junior • CEFR Mapping Procedure

04

Results

CEFR Score Mapping

05

Conclusion

1. BACKGROUND

What is CEFR?

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

A1–A2

Basic User

B1–B2

Independent User

C1–C2

Proficient User

Additional descriptors: 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

1. BACKGROUND

The CEFR Level Descriptors

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 fluently, flexibly, and effectively in academic or professional settings.

C2 Understands virtually everything heard or read. Can summarize and reformulate information, and express very precisely and naturally, even in demanding contexts.

1. BACKGROUND

Level-Setting in Testing

Purpose Classify test results using division (cut) scores, mapping the score scale onto CEFR's six standards and describing test-taker performance for each category.

Standard-Setting Process

- Balances and scales skill-description items
- Uses quantitative verification, such as the Rasch model
- Scaled scores are computed for cut scores based on summary statistics

Level-Setting Methods

Ebel, Nedelsky, and Angoff methods are common; the Angoff method is seen as balanced and practical. **Modified Angoff** is the most popular — it defines a standard minimum ability, facilitates panel discussion on judgments, and provides standard information based on past trial results.

Reference: Cizek & Bunch, 2011; Berk, 1986

Validity in Assigning Competency-Descriptive Items

Evaluation Factors of Level Setting (Cizek & Bunch, 2011)

Intrinsic	Extrinsic	Procedural
Consistency within and among panel members and with test-taker classifications	Comparing results with other criteria; verifying the appropriateness of cut scores using external data, like other tests	Clarity of purpose and process, ease of data analysis, panel training, systematic procedure, and panel confidence in results

Key Takeaway Validity is ensured through thorough processes, panelist evaluations, and external comparisons.

Overview of G-TELP Junior

Young Learners

Designed for elementary to middle school students

Two Skills

Tests include Listening and Reading sections

5 Levels

Organized into 5 levels tailored by age and difficulty

CEFR-Aligned

CEFR alignment enables international interpretation of scores

Test Structure by Level

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

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

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 G-TELP Korea and external researchers

Inspection Tool

- Analyzed G-TELP Junior test results from 2024
- Test takers: elementary, middle, and high school students
- Test components: listening comprehension and reading comprehension

Level-Setting Execution & Panelist Profile

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 a specific CEFR level

Panelist Profile (n = 10)

Category	Value
Gender	Male 60% • Female 40%
Nationality	USA 50% • S.Korea 30% • UK 10% • Pakistan/USA 10%
Mother Tongue	English 50% • Korean 30% • Eng/Chinese 10% • Pakistani 10%
Level-Setting Experience	Yes 20% • No 80%
Yrs. English Education	Avg 7.2 (SD 5.3, range 0–15)
Yrs. English Assessment	Avg 4.5 (SD 3.5, range 1–10)

Research Approach — Modified Angoff Method

Objective Validate split scores derived from the modified Angoff method.

Procedure

Modified Angoff: examinees' test sheets printed with correct answers.

Reasoning

Addresses large difficulty gaps, especially for tests with fewer easy or difficult questions.

Additional Info

For each item, panelists received the original item number, a 2/3 probability ability parameter, and the correct answer. Shared passages for the 1-passage-2-question type were indicated.

3. METHODOLOGY

CEFR Level-Setting Procedure

- Morning briefing: study's purpose, test details, and example questions
- Daily schedule: 9:00 AM – 5:00 PM, 4 rounds per day
- Panels divided by gender and background (in-house and external online groups)
- Activities: discuss and define minimum ability for CEFR levels
- Method: Modified Angoff — mark correct answers; 6 division points

Rounds 1 & 2

A2, B2, C2 Levels

Rounds 3 & 4

A1, B1, C1 Levels

4. RESULTS

CEFR Cut Scores by Test Level (Modified Angoff)

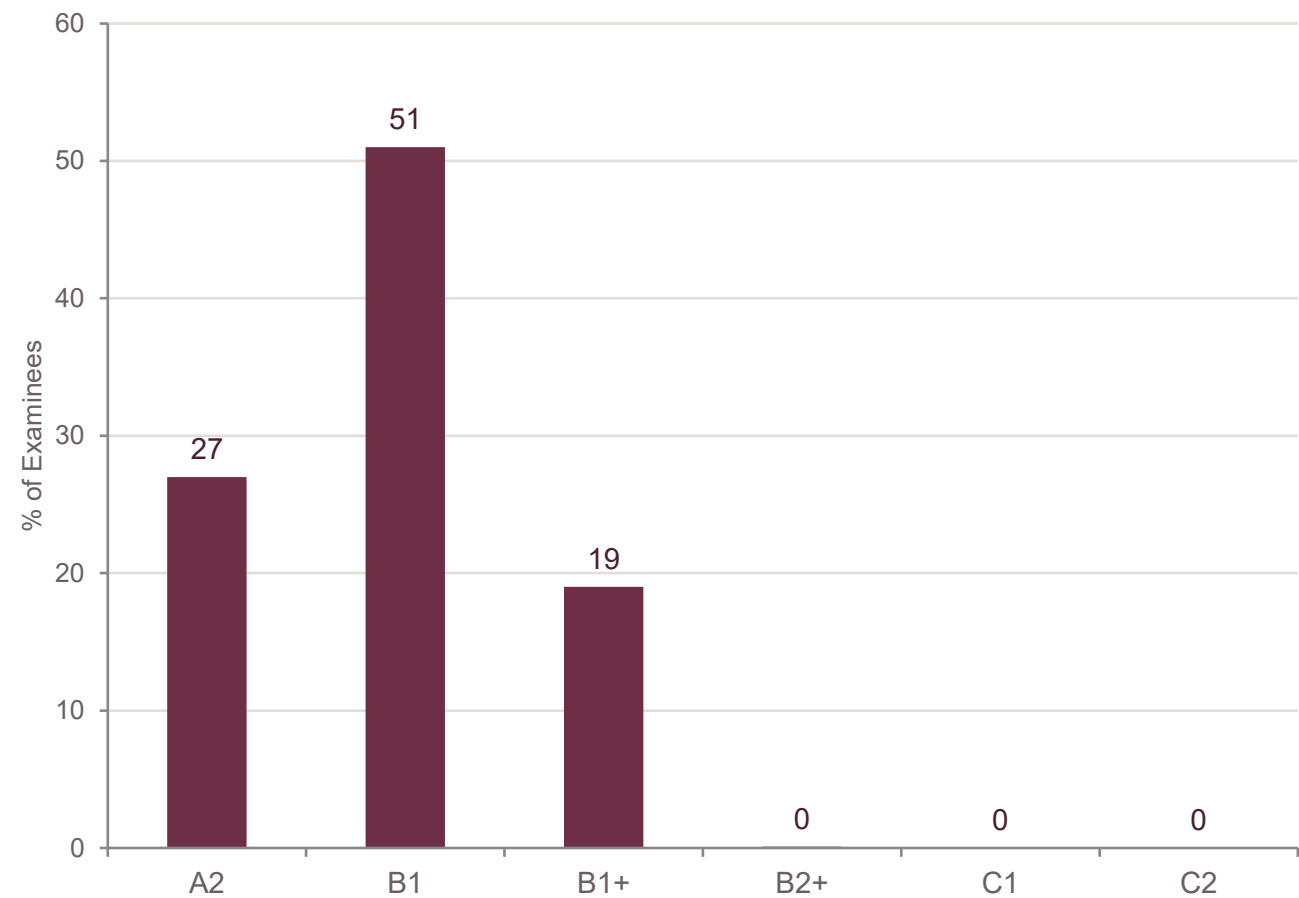
No outliers were detected (average used for calculations). Standard deviation across rounds 1–4 decreased, indicating convergence, and most levels saw a decreasing average difference with each round.

CEFR	Level 5	Level 4	Level 3	Level 2	Level 1
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

4. RESULTS

Test-Taker Distribution by CEFR Level

Assessing external validity of level-setting results by examining the distribution of examinees under the modified Angoff level setting.



By the Numbers

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

4. RESULTS

Survey — Level-Setting Process Evaluation

- Preliminary Data Evaluation: 99% found the provided preliminary data valuable
- Understanding of Study Purpose: 100% understood the purpose of the study
- Clarity of Instructions: all participants found the facilitators' instructions and explanations clear
- Feedback & Discussions: all participants found the feedback and discussions after each round useful

Statement	Strongly Agree	Agree
Homework assignment was useful preparation	50%	50%
Understood the purpose of the study	58%	42%
Facilitators' instructions/explanations were clear	63%	38%
Standard-setting training was adequate	75%	25%
Cut-score computation explanation was clear	25%	75%
Feedback/discussion between rounds was helpful	58%	42%
Standard-setting judgment process was easy to follow	63%	38%

Survey — Factors Influencing Level-Setting Judgments

- 1

Professional Experience

60% indicated this as a major influence
- 2

Personal Definition of Minimum Ability

Influenced 52% of participants
- 3

Group Discussions Between Rounds

Equally influenced 52% of participants
- 4

Split Scores from Other Panel Members

Relatively lower influence — 56% indicated impact

How influential was each information source on cutscore decisions?

Source	Very	Somewhat	Not
Definition of minimally competent person	52%	32%	16%
Between-round discussions	52%	32%	0%
Cutscores of other panel members	56%	24%	12%
Own professional experience	60%	40%	0%

Panel Confidence in the Final Cut Scores

High Confidence

- A1 — panel showed strong confidence in this level's final score
- C1 — similarly, a high degree of confidence was observed

Lower Confidence

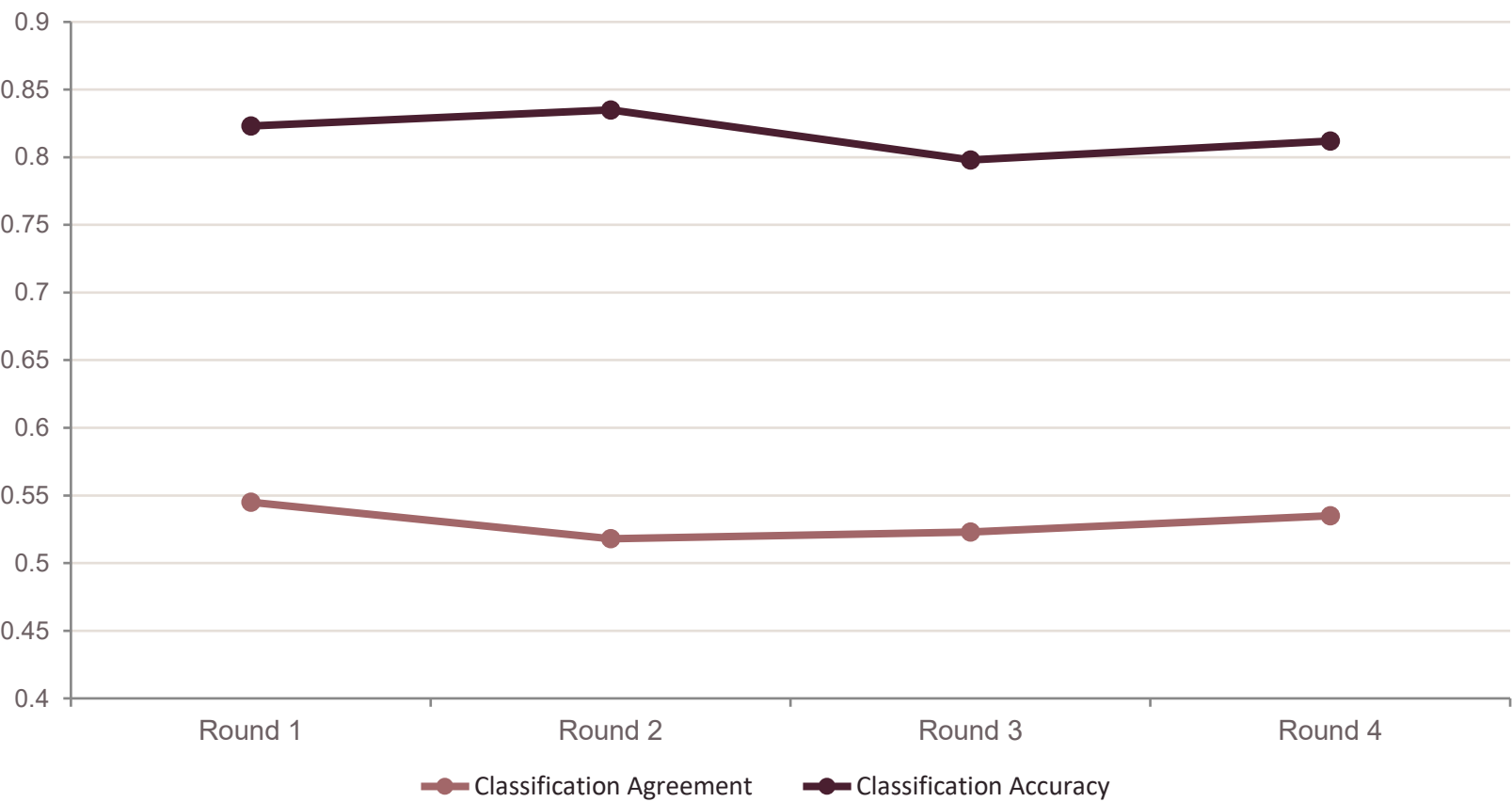
- A2 — panel had relatively less confidence in the division score
- B2 — confidence in the division score was also notably lower

CEFR Level	Very Comfortable	Somewhat Comfortable	Somewhat Unconf.	Very Unconf.
A1	29.17%	66.67%	4.17%	0.00%
A2	16.67%	70.83%	8.33%	0.00%
B1	20.83%	75.00%	4.17%	0.00%
B2	16.67%	70.83%	8.33%	0.00%
C1	29.17%	62.50%	8.33%	0.00%

4. RESULTS

Classification Agreement & Accuracy (Cohen's Kappa)

A slight dip was observed in Round 3, but recovery followed in subsequent rounds. As level numbers increased, both classification agreement and accuracy decreased, leading to slightly lower results in Round 3 versus Rounds 1–2.



Modified Angoff

Metric	R1	R2	R3	R4
Agreement	.545	.518	.523	.535
Accuracy	.823	.835	.798	.812

Reference: The Jamovi Project (2022), Version 2.3

CONCLUSION

Linking G-TELP Junior to CEFR: Methods & Validity

Objective Link the G-TELP Junior English Proficiency Test with CEFR.

Methodology

- Formation of an expert panel
- Utilized the modified Angoff method for cut-score calculation
- Ensured procedural standards for validity

5. CONCLUSION

Implications, Comparisons & Results

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 relative to expected European language standards.

Significance & Future Directions

Significance

- G-TELP Junior can discriminate across all six CEFR competency levels
- The 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+ and B2+)

Future Directions

- Link other G-TELP tests (Speaking, etc.) with CEFR
- Continue integrating CEFR across the G-TELP test family
- Explore CEFR level setting by comparing question difficulty
- Investigate equivalence of other English tests with G-TELP and the meaning of CEFR-linked scores

Thank You

Minjung Kim, PhD • ITSC G-TELP

A1

A2

B1

B2

C1

C2