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

Linking an English Language Speaking & Writing Test (GST & GWT) to the CEFR

2024

Prepared by
Minjung Kim, PhD

Conference
The Korea Association of Teachers of English International Conference

Date
July 5, 2024

Linking Levels of G-TELP Speaking Test (GST) and G-TELP Writing Test (GWT) to the Common European Framework of Reference for Languages (CEFR)

Abstract

This study aims to enhance understanding of the Common European Framework of Reference for Languages (CEFR) and provide a consistent and objective evaluation direction for English tests in Korea. To link the G-TELP Speaking Test (GST), Writing Test (GWT), and the Level 2 G-TELP English Proficiency Test to the CEFR, an expert panel was formed, and the benchmarking method was applied to derive proficiency levels.

The proficiency levels for the G-TELP Speaking and Writing Tests were calculated based on the six CEFR levels (A1, A2, B1, B2, C1, C2), defined by the CEFR. The validity of the evaluation was examined by considering internal, external, and procedural elements. The analysis of the collected scores showed a tendency for the standard deviation of the proficiency levels determined by the expert panel to decrease with each round. In addition to the six proficiency levels, detailed levels such as B1+ and B2+ were also derived. After determining the levels, most panel members expressed confidence in the final assessment scores. They also acknowledged that the preparation and process of the level-setting, as well as the explanations, were clear and helpful in their decision-making.

Keywords: level-setting, CEFR, benchmarking, G-TELP Speaking Test, G-TELP Writing Test

I. Purpose of research

The information derived regarding how well examinees achieve the performance standards in each area serves as fundamental data for formulating exam policies and plans in exams. For example, since 2003, the Korea Institute for Curriculum and Evaluation has set performance levels and linked scores to track academic achievement trends in national-level academic performance assessments.

To support the validity of test scores as criteria for distinguishing achievement levels, evidence showing the utility of the scores is needed. Messick's (1989) definition of validity and Kane's (1992, 2006, 2013) argument-based approach to validation emphasize the interpretation and use of test results. When large-scale data collection is practically challenging, linking scores to international frameworks like the CEFR and indirectly comparing the meaning of scores between tests may serve as an alternative.

The CEFR is widely used as a standard measure of foreign language proficiency. It helps people learn foreign languages independently and provides consistent content and methods to language education institutions, enabling objective evaluation. Moreover, the CEFR allows for comparative analysis with English tests developed abroad, increasing the usability of scores. When large-scale data collection for score linking between tests is challenging, linking to international standards like the CEFR and indirectly comparing the meaning of scores can serve as a viable alternative.

Most English proficiency tests have studied their alignment with the CEFR. In 2008, the Educational Testing Service (ETS) published the results of correlating the TOEFL iBT and

TOEIC with the CEFR (Tannenbaum & Wylie, 2008). They used the CEFR alignment guidelines to match the TOEFL iBT and TOEIC speaking, writing, listening, and reading sections with the CEFR. Several studies on the correlation between the International English Language Testing System (IELTS) and the CEFR have also been conducted, with the latest level-setting study published in 2013 (Lim, Geranpayeh, Khalifa, & Buckendahl, 2013).

In South Korea, numerous studies have validated the achievement standards of the national curriculum by comparing them with the internationally verified CEFR. The CEFR has also been compared with the achievement standards for first-year high school English. When developing achievement standards, it is argued that activities, examples, and evaluation samples should be provided to make the standards actionable (Lee & Kim, 2009). Hwang (2016) claimed that the CEFR is a global standard in foreign language education, allowing for the diagnosis and evaluation of language proficiency based on international criteria, making it highly efficient in education.

The purpose of this study is to link the G-TELP Speaking Test (GST) and G-TELP Writing Test (GWT) with the CEFR. To match the CEFR scale from A1 to C2, the benchmarking method was improved and applied to derive proficiency levels, and the validity was examined based on 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 as a standard tool for measuring foreign language proficiency in education. In the 1990s, both intuitive, qualitative, and quantitative methods were used in the process of developing the CEFR levels and their descriptors. The CEFR provides metalinguistic reference points for foreign language teachers, learners, assessors, and textbook developers, allowing language education institution leaders, textbook authors, teachers, teacher trainers, and test developers to reflect on their efforts and whether they meet learners' needs.

The CEFR consists of six proficiency levels (A1, A2, B1, B2, C1, C2), each with detailed achievement descriptors for speaking, writing, listening, and reading skills and subskills. In addition to these six levels, descriptors for levels such as Pre-A1, A2+, B1+, and B2+ are also included. A1 and A2 represent basic levels, B1 and B2 represent independent levels, and C1 and C2 represent proficient levels.

By adopting the CEFR, educational objectives based on objectivity, transparency, and commonality can be established, and global and universal standards can be set in foreign language proficiency tests. The CEFR contains numerous scales, including subscales for speaking, writing, listening, and reading, as well as communication strategies, linguistic, sociolinguistic, and pragmatic competencies. It avoids judging language proficiency solely based

on overall scores and instead provides detailed task-specific descriptors, categorized into three areas: communicative activities, communicative strategies, and communicative competence. Additionally, the CEFR divides language activities into private, public, occupational, and educational domains, describing language use in various contexts (Kim, 2019).

<Table 1> Descriptors of the Six CEFR Levels

A1	The lowest level, where learners can understand and use familiar, everyday expressions and basic phrases. For example, learners can introduce themselves and ask questions about personal details, but only if the other person speaks slowly and clearly.
A2	At this level, learners can understand and use phrases and expressions related to familiar topics such as family, shopping, and local geography. They can exchange information in simple terms about their background and surroundings.
B1	Learners can understand the main points of clear, standard speech or writing on familiar matters related to work, school, or leisure. They can handle most situations while traveling in a region where the language is spoken and describe experiences, events, dreams, and opinions.
B2	Learners can understand the main ideas of complex text on concrete and abstract topics, including technical discussions in their field of specialization. They can interact fluently with native speakers without strain and express themselves on a wide range of topics.
C1	At this level, learners can understand a wide range of longer and more demanding texts and recognize implicit meaning. They can express themselves fluently and spontaneously for social, academic, and professional purposes and produce clear, well-structured texts on complex subjects.
C2	Learners can effortlessly understand virtually everything they hear or read. They can summarize and synthesize information from different sources and express themselves spontaneously with high precision, even in complex situations.

2. Level-Setting

Level-setting refers to the process of determining one or more cut scores in an exam. Using these cut scores, examinees' performance is categorized into two or more levels (Cizek & Bunch, 2011). Cut scores are determined through standard-setting procedures, and performance descriptors are developed for each level. In this study, the levels were classified into six categories based on the CEFR scale (A1, A2, B1, B2, C1, C2), with additional classifications for levels like B1+ and B2+.

In language testing, the benchmarking method is suitable for directly testing speaking and writing skills, as it involves a more natural procedure compared to other standard-setting methods. Benchmarking allows expert panels to establish achievement standards by comparing local test tasks with CEFR levels and developing a shared understanding of those levels. Through this process, proficiency levels for the G-TELP Speaking and Writing Tests were established.

A. Expert Panel Composition

The panel consisted of 10 members, including learners, measurement experts, textbook authors, educators, and instructors. The panel included 4 native speakers from the internal global research team, 2 internal researchers, 1 internal textbook researcher, 1 native English professor, 1 native English instructor, and 1 internal/external native researcher. The panel members were sent materials related to the CEFR, level-setting, and schedules two weeks before the actual level-setting process. The materials sent included the following:

- Common European Framework of Reference for Languages (CEFR), Council of Europe (2001)
- Companion Volume to the CEFR, Council of Europe (2018)
- Manual for Relating Language Examinations to the CEFR, Council of Europe (2009)
- CEFR scales related to Speaking and Writing
- G-TELP Speaking and Writing test items linked to the CEFR scales
- Research linking TOEFL and TOEIC to the CEFR by ETS (Tannenbaum & Wylie, 2008)
- Background variables questionnaire
- Survey questionnaire
- CEFR level samples for Speaking and Writing items
- CEFR level descriptors for Speaking and Writing
- Schedule for the level-setting sessions

Before participating in the level-setting, the panel members reviewed the materials to familiarize themselves with the CEFR and the level-setting process. The CEFR provides detailed descriptors for various sub-areas of language ability. To aid in the panel's understanding, only the reading-related descriptors were compiled and provided. Additionally, a collection of reading items developed or aligned with the CEFR by other experts was provided to help panel members compare their understanding of the CEFR scales with their own judgments.

The level-setting process was conducted over three weeks, starting from the third week of December 2022, across three days. On the first day, the panel members gathered and were briefed on the research objectives, the purpose of the exam, test details, and sample items. The panel members were divided into three groups, considering their gender and background, to

understand the CEFR and level-setting methods through group activities. The panel members received CEFR scales and discussed the level-setting method to adapt to it and understand the concepts. For each area, the minimum competence required at each CEFR level was defined. After group activities, the definitions of minimum competence provided by each group were shared and discussed among the entire panel.

<Table 2> 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

In this study, considering the relatively small number of test items in each area, level-specific scores for each test were derived using the benchmarking method. The panel members were provided with a variety of feedback materials to examine the test from multiple perspectives and

to facilitate group discussions. Panel members were also instructed to mark items as "Not Applicable (N/A)" if they deemed that certain test items were not suitable for examinees at specific CEFR levels. If one-third or more of the panel marked an item as N/A, no cut score was calculated for that level.

B. Procedure

The Speaking and Writing tests were direct tests, and the level-setting process was conducted using the benchmarking method, as it involved a direct evaluation of proficiency levels (in this case, the six CEFR levels). The benchmarking process provided samples that demonstrated performance at each level, allowing the panel members to understand the CEFR levels from A1 to C2. The process began with sample analysis and evaluation, following examples of the CEFR benchmarks from A1 to C2.

The benchmarking process was conducted over three days. On the first morning, the panel members gathered for an orientation on the research objectives, test purposes, detailed test items, and sample items. The daily schedule ran from 9 AM to 5 PM, with each area being assessed over 1–2 rounds.

- Day 1: Orientation, GST (Speaking), and benchmarking with discussion
- Day 2: GST (Speaking), first and second rounds of each level with discussion
- Day 3: GWT (Writing), first and second rounds of each level with discussion

To facilitate discussions, the panel was divided into two groups based on gender and background: one group within the company and one external online group. Each group engaged in activities to understand the CEFR.

Through group discussions, the panel defined the minimum competencies for each CEFR level. These competencies were then shared with the entire panel, allowing everyone to internalize the concepts and adapt to the level-setting method. The panel members reviewed various feedback materials, assessed the test from different angles, and engaged in group discussions to make the process more comprehensive. To apply the benchmarking method, the test papers were printed out, and for the speaking test, the correct answer recordings were played for the panel to assess. For the writing test, the correct answers were marked. The six CEFR levels were considered, and the panel members were instructed to evaluate examinees from the lowest level (A1) to the highest (C2). If scores were unexpectedly high or low, panel members revisited previous answers and analyzed them based on the rubric for a more detailed evaluation.

In the first round, panel members were given enough time to review the test papers. For the benchmarking method, panel members were instructed to determine the minimum competence required for each test item from the first round. In the second round, the panel members either maintained or adjusted the level of each item based on the first round's discussions. In the first round, panel members made independent judgments without group discussions. After the round, feedback materials were provided, and panel members took a break. After enough group discussion, the next round began. The tests used in the research were analyzed based on actual examinee performance, allowing feedback materials to be prepared. A

unique number was assigned to each panel member to avoid any pressure to modify their cut scores during group discussions, ensuring that feedback was provided anonymously.

When evaluating samples, the process should follow logical steps, including reaching a consensus, presenting illustrations, conducting individual assessments, small group evaluations, and full group discussions, with graphical data collection and feedback provision wherever possible. After training on the test tasks, the coordinator ensures that all necessary materials are available to the panel members before the benchmarking/standard-setting process begins.

During the sessions, the coordinator must summarize opinions and discussions to achieve reliable results through the most appropriate judgments. It is important to remember that participants are asked to estimate the level, and the group is not asked to form a consensus on the sample levels. Instead, predetermined criteria are applied to reach the correct level, regardless of previous evidence. Before becoming accustomed to the training, judgments should not be too strict or lenient, as this may destabilize the panel members' future judgments. Therefore, investing sufficient time in pre-training is essential.

Coordinators should ensure that participants become familiar with standardized samples, understanding why and how a specific sample corresponds to a certain level. Working in pairs or small groups is usually well-received by participants. If necessary, the coordinator can facilitate the discussions, guiding the group in the right direction. The primary benefit of this group work is that participants are naturally compelled to justify their judgments by using clearly defined criteria.

In speaking evaluations, panel members must have a proficiency level of at least B2/C1 and begin by analyzing and evaluating CEFR benchmark performances. Most audio samples follow a similar format, including a monologue phase (one candidate explaining something to another candidate) and an interaction phase (two candidates engaging in a discussion). For writing assessments, it is also important to review samples of written interaction, such as memos, letters, and written compositions (e.g., explanations, stories, reviews).

C. Statistical Analysis of Local Samples Subject to Benchmarking

The grades of local samples subject to benchmarking need to be statistically analyzed. (a) Confirm the relationship with levels, (b) calculate intra-rater reliability (consistency) and inter-rater reliability (agreement among participants), and (c) evaluate the degree of agreement among participants. The average level is derived by analyzing the grades during the benchmarking process. The main advantage is that it allows identifying inconsistent panelists and excluding them if necessary.

At the end of the session, the set of benchmarked samples and detailed records kept during the session will be very helpful for future training. Documentation for each sample, specifying which level the sample represents, can act as a model. Audio recordings of discussions during the session can serve as a useful resource for preparing such discussions.

<Table 3> Oral Assessment Criteria GRID (CEFR)

	RANGE	ACCURACY	FLUENCY	INTERACTION	COHERENCE
C2	Shows great flexibility reformulating ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis, to differentiate and to eliminate ambiguity. Also has a good command of idiomatic expressions and colloquialisms.	Maintains consistent grammatical control of complex language, even while attention is otherwise engaged (e.g. in forward planning, in monitoring others' reactions).	Can express him/herself spontaneously at length with a natural colloquial flow, avoiding or backtracking around any difficulty so smoothly that the interlocutor is hardly aware of it.	Can interact with ease and skill, picking up and using non-verbal and intonational cues apparently effortlessly. Can interweave his/her contribution into the joint discourse with fully natural turntaking, referencing, allusion making etc.	Can create coherent and cohesive discourse making full and appropriate use of a variety of organisational patterns and a wide range of connectors and other cohesive devices.
C1	Has a good command of a broad range of language allowing him/her to select a formulation to express him/herself clearly in an appropriate style on a wide range of general, academic, professional or leisure topics without having to restrict what he/she wants to say.	Consistently maintains a high degree of grammatical accuracy; errors are rare, difficult to spot and generally corrected when they do occur.	Can express him/herself fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language.	Can select a suitable phrase from a readily available range of discourse functions to preface his remarks in order to get or to keep the floor and to relate his/her own contributions skilfully to those of other speakers.	Can produce clear, smoothly flowing, well-structured speech, showing controlled use of organisational patterns, connectors and cohesive devices.
B2+					
B2	Has a sufficient range of language to be able to give clear descriptions, express viewpoints on most general topics, without much conspicuous searching for words, using some complex sentence forms to do so.	Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstanding, and can correct most of his/her mistakes.	Can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he or she searches for patterns and expressions, there are few noticeably long pauses.	Can initiate discourse, take his/her turn when appropriate and end conversation when he/she needs to, though he/she may not always do this elegantly. Can help the discussion along on familiar ground confirming comprehension, inviting others in, etc.	Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse, though there may be some "jumpiness" in a long contribution.
B1+					
B1	Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocutions on topics such as family, hobbies and interests, work, travel, and current events.	Uses reasonably accurately a repertoire of frequently used "routines" and patterns associated with more predictable situations.	Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production.	Can initiate, maintain and close simple face-to-face conversation on topics that are familiar or of personal interest. Can repeat back part of what someone has said to confirm mutual understanding.	Can link a series of shorter, discrete simple elements into a connected, linear sequence of points.
A2+					
A2	Uses basic sentence patterns with memorised phrases, groups of a few words and formulae in order to communicate limited information in simple everyday situations.	Uses some simple structures correctly, but still systematically makes basic mistakes.	Can make him/herself understood in very short utterances, even though pauses, false starts and reformulation are very evident.	Can ask and answer questions and respond to simple statements. Can indicate when he/she is following but is rarely able to understand enough to keep conversation going of his/her own accord.	Can link groups of words with simple connectors like "and, "but" and "because".
A1	Has a very basic repertoire of words and simple phrases related to personal details and particular concrete situations.	Shows only limited control of a few simple grammatical structures and sentence patterns in a memorised repertoire.	Can manage very short, isolated, mainly pre-packaged utterances, with much pausing to search for expressions, to articulate less familiar words, and to repair communication.	Can ask and answer questions about personal details. Can interact in a simple way but communication is totally dependent on repetition, rephrasing and repair.	Can link words or groups of words with very basic linear connectors like "and" or "then".

<Table 4> Suppleentary Criteria GRID: “Pluse Levels”(CEFR)

	RANGE	ACCURACY	FLUENCY	INTERACTION	COHERENCE
C2					
C1					
B2+	Can express him/herself clearly and without much sign of having to restrict what he/she wants to say.	Shows good grammatical control; occasional “slips” or non-systematic errors and minor flaws in sentence structure may still occur, but they are rare and can often be corrected in retrospect.	Can communicate spontaneously, often showing remarkable fluency and ease of expression in even longer complex stretches of speech. Can use circumlocution and paraphrase to cover gaps in vocabulary and structure.	Can intervene appropriately in discussion, exploiting a variety of suitable language to do so, and relating his/her own contribution to those of other speakers.	Can use a variety of linking words efficiently to mark clearly the relationships between ideas.
B2					
B1+	Has a sufficient range of language to describe unpredictable situations, explain the main points in an idea or problem with reasonable precision and express thoughts on abstract or cultural topics such as music and films.	Communicates with reasonable accuracy in familiar contexts; generally good control though with noticeable mother tongue influences.	Can express him/herself with relative ease. Despite some problems with formulation resulting in pauses and “cul-de-sacs”, he/she is able to keep going effectively without help.	Can exploit a basic repertoire of strategies to keep a conversation or discussion going. Can give brief comments on others’ views during discussion. Can intervene to check and confirm detailed information.	*No descriptor available*
B1					
A2+	Has sufficient vocabulary to conduct routine, everyday transactions involving familiar situations and topics, though he/she will generally have to compromise the message and search for words.	*No descriptor available*	Can adapt rehearsed memorised simple phrases to particular situations with sufficient ease to handle short routine exchanges without undue effort, despite very noticeable hesitation and false starts.	Can initiate, maintain and close simple, restricted face-to-face conversation, asking and answering questions on topics of interest, pastimes and past activities. Can interact with reasonable ease in structured situations, given some help, but participation in open discussion is fairly restricted.	Can use the most frequently occurring connectors to link simple sentences in order to tell a story or describe something as a simple list of points.
A2					
A1					

<Table 5> Written Assessment Criteria GRID

	Overall	Range	Coherence	Accuracy	Description	Argument
C2	Can write clear, *highly accurate and* smoothly flowing complex texts in an appropriate and effective *personal* style *conveying finer shades of meaning. Can use* a logical structure which helps the reader to find significant points.	Shows great flexibility in *formulating* ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis and to eliminate ambiguity. Also has a good command of idiomatic expressions and colloquialisms.	Can create coherent and cohesive texts making full and appropriate use of a variety of organisational patterns and a wide range of connectors and other cohesive devices.	Maintains consistent *and highly accurate* grammatical control of *even the most complex language forms. Errors are rare and concern rarely used forms.*	Can write clear, smoothly flowing and fully engrossing stories and descriptions of experience in a style appropriate to the genre adopted.	Can produce clear, smoothly flowing, complex reports, articles and essays which present a case or give critical appreciation of proposals or literary works. Can provide an appropriate and effective logical structure which helps the reader to find significant points.
C1	Can write clear, well-structured *and mostly accurate* texts of complex subjects. *Can underline* the relevant salient issues, *expand and support* points of view at some length with subsidiary points, reasons and relevant examples, and *round* off with an appropriate conclusion.	Has a good command of a broad range of language allowing him/her to select a formulation to express him/herself clearly in an appropriate style on a wide range of general, academic, professional or leisure topics without having to restrict what he/she wants to say. *The flexibility in style and tone is somewhat limited.*	Can produce clear, smoothly flowing, well-structured text, showing controlled use of organisational patterns, connectors and cohesive devices.	Consistently maintains a high degree of grammatical accuracy; *occasional errors in grammar, collocations and idioms.*	Can write clear, detailed, well-structured and developed descriptions and imaginative texts in a mostly assured, personal, natural style appropriate to the reader in mind.	Can write clear, well-structured expositions of complex subjects, underlining the relevant salient issues. Can expand and support point of view with some subsidiary points, reasons and examples.
B2	Can write clear, detailed *official and semi-official* texts on a variety of subjects related to his field of interest, synthesising and evaluating information and arguments from a number of sources. *Can make a distinction between formal and informal language with occasional less appropriate expressions.*	Has a sufficient range of language to be able to give clear descriptions, express viewpoints on most general topics, using some complex sentence forms to do so. *Language lacks, however, expressiveness and idiomaticity and use of more complex forms is still stereotypic.*	Can use a number of cohesive devices to link his/her sentences into clear, coherent text, though there may be some "jumpiness" in a longer text.	Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstandings.	Can write clear, detailed descriptions of real or imaginary events and experiences marking the relationship between ideas in clear connected text, and following established conventions of the genre concerned. Can write clear, detailed descriptions on a variety of subjects related to his/her field of interest. Can write a review of a film, book or play.	Can write an essay or report that develops an argument systematically with appropriate highlighting of some significant points and relevant supporting detail. Can evaluate different ideas or solutions to a problem. Can write an essay or report which develops an argument, giving some reasons in support of or against a particular point of view and explaining the advantages and disadvantages of various options. Can synthesise information and arguments from a number of sources.
B1	Can write straightforward connected texts on a range of familiar subjects within his field of interest, by linking a series of shorter discrete elements into a linear sequence. *The texts are understandable but occasional unclear expressions and/or inconsistencies may cause a break-up in reading.*	Has enough language to get by, with sufficient vocabulary to express him/herself with some circumlocutions on topics such as family, hobbies and interests, work, travel, and current events.	Can link a series of shorter discrete elements into a connected, linear text.	Uses reasonably accurately a repertoire of frequently used "routines" and patterns associated with more *common* situations. *Occasionally makes errors that the reader usually can interpret correctly on the basis of the context.*	Can write accounts of experiences, describing feelings and reactions in simple connected text. Can write a description of an event, a recent trip – real or imagined. Can narrate a story. Can write straightforward, detailed descriptions on a range of familiar subjects within his field of interest.	Can write short, simple essays on topics of interest. Can summarise, report and give his/her opinion about accumulated factual information on a familiar routine and non-routine matters, within his field with some confidence. Can write very brief reports to a standard conventionalised format, which pass on routine factual information and state reasons for actions.
A2	Can write a series of simple phrases and sentences linked with simple connectors like "and", "but" and "because". *Longer texts may contain expressions and show coherence problems which makes the text hard to understand.*	Uses basic sentence patterns with memorized phrases, groups of a few words and formulae in order to communicate limited information mainly in everyday situations.	Can link groups of words with simple connectors like "and", "but" and "because".	Uses simple structures correctly, but still systematically makes basic mistakes. *Errors may sometimes cause misunderstandings.*	Can write very short, basic descriptions of events, past activities and personal experiences Can write short simple imaginary biographies and simple poems about people.	
A1	Can write simple isolated phrases and sentences. *Longer texts contain expressions and show coherence problems which make the text very hard or impossible to understand.*	Has a very basic repertoire of words and simple phrases related to personal details and particular concrete situations.	Can link words or groups of words with very basic linear connectors like "and" and "then".	Shows only limited control of a few simple grammatical structures and sentence patterns in a memorized repertoire. *Errors may cause misunderstandings.*	Can write simple phrases and sentences about themselves and imaginary people, where they live and what they do, etc.	

3. Validation Methods

The standard deviation of the proficiency levels proposed by each panel member was examined, and Cohen's Kappa coefficient and classification accuracy coefficients were calculated to analyze the consistency and accuracy of classification based on final proficiency levels as an internal criterion for validating the level-setting process (Hanson & Brennan, 1990; Lee, Hanson & Brennan, 2002). The classification accuracy coefficient represents how accurately observed scores classify test-takers when compared to their true scores. It is the difference between the false positive rate (the probability that a test-taker whose true score τ is below the true level $\tau 0$ obtains an observed score x0 above their actual level) and the false negative rate (the probability that a test-taker whose true score τ is above $\tau 0$ obtains an observed score x0 below their actual level) (Hanson & Brennan, 1990).

In this study, classification consistency and accuracy were calculated using item response theory (Hanson & Brennan, 1990; Lee, Hanson & Brennan, 2002). Specifically, using item and ability parameter distributions estimated from actual data, true score distributions and error score distributions were obtained, and Cohen's Kappa and classification accuracy coefficients were calculated based on the proficiency levels derived in this study. The Jamovi 2.3.2 software (Lee & Kolen, 2008) was used to calculate these coefficients. As an external validity criterion, the distribution of test-takers proficiency levels based on CEFR levels was calculated. Lastly, procedural validity was secured by reviewing the final proficiency levels after completing two rounds and administering a survey to evaluate the level-setting process. The survey included questions about the evaluation of the level-setting process, factors influencing level-setting decisions, and confidence in the final proficiency levels.

IV. Results

1. G-TELP Speaking Test Level-Setting Results

The results of the level-setting for the G-TELP Speaking Test (GST) are shown in **Table 6**. The levels calculated through the benchmarking method are indicated in bold after rounds 1 and 2 of benchmarking, covering levels A1, A2, B1, B2, C1, and C2.

During the final stage of deriving the proficiency levels, all 10 panel members proposed proficiency levels for all CEFR levels, without marking any items as "N/A." There were no extreme values, so the average was used without considering other statistical measures. To verify internal validity, the standard deviation proposed by each panel member was examined. **Table 6** shows the results of level-setting through the benchmarking method. During rounds 1 and 2, the standard deviation between the levels determined by each panel member decreased and showed a tendency to converge in a consistent direction. In most levels, the difference in averages decreased as the rounds progressed.

Table 6 summarizes the results of the benchmarking level-setting process, comparing the results of GST levels to CEFR levels, and also comparing levels of G-TELP Level 2, OPIc, and TOEIC Speaking to the benchmarking results for GST.

<Table 6> Comparison of GST and Other Tests by CEFR Levels: Results of Level Setting Based on Benchmarking Methods

CEFR	GST	GTELP Level 2	OPIc	TOEIC Speaking
A1	**21-25**	**Level 9, 10**	Novice High	60-80
A2	**26-43**	**Level 8**	Intermediate Low	90-100
B1	**44-55(B1)** **56-73(B1+)**	**Level 7** **Level 6** **Level 5** **Level 4**	Intermediate Mid 1 Intermediate Mid 2 Intermediate Mid 3 Intermediate High	100-130 140-150
B2	**74-85(B2)** **86-94(B2+)**	**Level 3**	Advanced Low Advanced Mid	160-170 180-190
C1	**95-99**	**Level 2**	Advanced High	200
C2		**Level 1**	Superior	

Table 7 summarizes the survey questions related to the level-setting of GST. The analysis showed that 99% of the respondents found the preparatory materials provided before the level-setting to be useful, and 100% responded that they understood the purpose of the study. Additionally, all respondents evaluated that the instructions and explanations provided by the facilitator were clear, the explanations of the level-setting method were detailed, and the explanation of the grade calculation method was clear. They also found the feedback and

discussions provided after each round of level-setting to be useful and reported that the process of making level-setting judgments was easy to follow.

<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.	33%	56%		
I understood the purpose of the study.	67%	33%		
The instructions and explanations provided by the facilitators were clear.	567%	33%		
The training in the standard-setting methods was adequate to give me the information I needed to complete my assignment.	33%	56%	11%	
The explanation of how the recommended cut scores were computed was clear.	44%	56%		
The opportunity for feedback and discussion between rounds was helpful.	67%	33%		
The process of making the standard-setting judgments was easy to follow.	33%	67%		

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

Table 8 summarizes the factors that influenced level-setting judgments in the GST survey. The most influential factors were group discussions between rounds (67%), the definition of the minimally competent person (56%), and the respondents' own professional experience

(56%). On the other hand, the influence of other panel members' proposed levels was reported to be relatively low (33%).

<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	56%	33%	11%
The between-round discussions*	67%	11%	11%
The cutscores of other panel members	33%	56%	0%
My own professional experience*	56%	44%	0%

Table 9 shows the survey results regarding the panel members' confidence in the final cut scores for each CEFR level. The panel members who participated in setting the levels for G-TELP Level 2 expressed high confidence in the C1 and B2 levels, but relatively lower confidence in the A1 and A2 levels.

<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	11%	44%	33%	11%
Cutscore for CEFR A2	11%	78%	11%	0.00%
Cut score for CEFR B1	11%	89%	0%	0.00%
Cut score for CEFR B2	22%	67%	11%	0.00%
Cut score for CEFR C1	22%	67%	11%	0.00%

Due to missing data and multiple responses, the total may appear to be less than or greater than 100%.

Table 10 presents the classification consistency and classification accuracy after each round, which serve as internal criteria for evaluating the validity of the level-setting process. The Kappa coefficient was used to measure classification consistency. The analysis showed that both classification consistency and accuracy slightly increased from round 1 to round 2. In the benchmarking method, both classification consistency and accuracy improved as the process progressed from the first to the second round. Considering the number of levels classified in this study and the range observed in previous research, the classification consistency and accuracy were found to be high.

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

		Round 1	Round 2
Modified Angoff	Classification Agreement	0.545	0.551
	Classification Accuracy	0.792	0.793

2. G-TELP Writing Test Level-Setting Results

The level-setting results for the G-TELP Level 2 exam are shown in **Table 4**. The levels calculated through the benchmarking method are indicated in bold for levels A1, A2, B1, B2, C1, and C2, as determined through rounds 1 and 2 of the benchmarking process. During the final proficiency level derivation, all 10-panel members proposed cut scores for all CEFR levels without marking any items as "N/A." Additionally, since there were no extreme values, the average was used without considering other statistical measures. To verify internal validity, the standard deviation proposed by each panel member was examined.

Table 11 shows the level-setting results using the benchmarking method. Over rounds 1 and 2, the standard deviation between the cut scores proposed by each panel decreased, showing a tendency to converge in a consistent direction. As the rounds progressed, the difference in averages generally decreased across most levels. The A2 and B2 level scores tended to decrease, while the B2+ level showed a slight increase. **Table 11** provides a comparison of the level-setting results based on the final round, organized into a CEFR comparison table, showing the benchmarking results for GWT, OPIc Writing, and TOEIC Writing.

<Table 11> Comparison of GWT and Other Tests by CEFR Levels: Results of Level Setting Based on Benchmarking Methods

CEFR	GWT	OPIc Writing	TOEIC Writing
A1	**Level 8**	Novice High	50-60
A2	**Level 7**	Intermediate Low	70-80
B1	**Level 6**	Intermediate Mid 1 Intermediate Mid 2 Intermediate Mid 3 Intermediate High	90-100 110-120
B2	**Level 5** **Level 4**	Advanced Low	130-150
C1	**Level 3** **Level 2**	Advanced Mid Advanced High	160-170 180-200
C2	**Level 1**	Superior	

Table 12 summarizes the survey questions related to the level-setting of GWT.

The analysis showed that 99% of the respondents found the preparatory materials provided before participating in the level-setting to be useful, and 100% responded that they understood the purpose of the study. Additionally, all respondents evaluated that the instructions and explanations provided by the facilitator were clear, the explanations of the level-setting method were detailed, and the explanation of the grade calculation method was clear. The feedback and discussions provided after each round of level-setting were also found to be useful, and the respondents reported that the process of making level-setting judgments was easy to follow. However, some respondents felt that the training on the level-setting method was not entirely adequate for completing the tasks.

<Table 12> GWT 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.	33%	56%		
I understood the purpose of the study.	44%	56%		
The instructions and explanations provided by the facilitators were clear.	56%	44%		
The training in the standard-setting methods was adequate to give me the information I needed to complete my assignment.	22%	56%	11%	
The explanation of how the recommended cut scores were computed was clear.	22%	78%		
The opportunity for feedback and discussion between rounds was helpful.	56%	44%		
The process of making the standard-setting judgments was easy to follow.	33%	67%		

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

Table 13 summarizes the factors that influenced level-setting judgments from the survey questions. The most influential factors were the respondents' own professional experience (67%), group discussions between rounds (56%), and the definition of the minimally competent person (44%). On the other hand, the influence of other panel members' proposed levels was reported to be relatively low (11%).

<Table13> 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	44%	33%	22%
The between-round discussions*	56%	11%	22%
The cutscores of other panel members	11%	56%	22%
My own professional experience*	67%	33%	0%

Table 14 shows the survey results regarding the panel members' confidence in the final cut scores for each CEFR level. The panel members who participated in the GWT level-setting process expressed high confidence in the cut scores for C1, A1, A2, and B2, but showed relatively lower confidence in the cut score for B1.

<Table 14> 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	11%	89%	11%	0.00%
Cutscore for CEFR A2	11%	89%	11%	0.00%
Cut score for CEFR B1	11%	78%	22%	0.00%
Cut score for CEFR B2	11%	89%	11%	0.00%
Cut score for CEFR C1	22%	78%	11%	0.00%

Table 15 presents the classification consistency and classification accuracy of the cut scores derived after each round, which serve as internal criteria for evaluating the validity of the

level-setting process. The Kappa coefficient was used to measure classification consistency. The analysis showed that both classification consistency and accuracy for the six proficiency levels (A1, A2, B1, B2, C1, C2) slightly increased from round 1 to round 2. In the benchmarking method, both classification consistency and accuracy continued to improve from round 1 to round 2. Although classification consistency and accuracy tend to decrease as the number of levels increases, the results still showed higher values compared to round 1. Considering the range observed in previous studies and the number of levels classified in this study, the consistency and accuracy of the classifications were found to be high.

<Table 15> Classification Agreement and Classification Accuracy Coefficients for Split Scores in Each Round

		Round 1	Round 2
Modified Angoff	Classification Agreement	0.554	0.581
	Classification Accuracy	0.812	0.830

V. Conclusions

In this study, a panel of experts was formed, and the benchmarking method was used to align the G-TELP English proficiency test by ITSC in the U.S. with the CEFR scale. As a result, cut scores for all areas of the G-TELP test were determined according to the six CEFR levels. The internal, external, and procedural validity criteria supporting the validity of this level-setting study were secured as follows. Internally, the standard deviation of the cut scores determined by the expert panel members decreased over successive rounds. The classification consistency and accuracy coefficients of proficiency levels also showed good values, validating the proficiency levels. Externally, the distribution of test-takers across the CEFR levels was appropriate based on the level-setting results. Procedurally, the characteristics of the panel and the entire level-setting process were documented, and most panel members evaluated that the preparation, explanation, and guidance during the level-setting process were clear and helpful for making decisions. Additionally, the majority of panel members expressed confidence in the final proficiency levels.

The difficulty of linking test scores to the CEFR should not be underestimated. According to Weir (2005), the CEFR does not provide sufficient information on how situational factors affect performance or how language develops across levels. Milanovic (2009) noted that "the CEFR is deliberately underspecified and incomplete" (p. 3), emphasizing that the CEFR is meant to describe the characteristics of levels rather than define them precisely. There can be difficulties in interpreting the differences across CEFR levels consistently (Alderson et al., 2006; Papageorgiou, S., 2010). Some of these difficulties were clearly revealed in panel discussions. When developing qualified explanations, CEFR panels discover that the descriptive language is not applied consistently across the entire range of CEFR levels.

However, the difficulty of alignment also depends on the nature of the test. The G-TELP Speaking and Writing tests align well with the CEFR, and tests specifically developed for CEFR mapping tend to encounter fewer alignment issues than those that were not designed for this purpose.

In this study, while the tests measured two major language skills, both were covered by the CEFR, and the test items and tasks were not specially developed for alignment purposes. These skills are described by the CEFR, and indeed, the G-TELP Speaking and Writing tests existed before the CEFR. All target CEFR levels were mapped, and there was positive procedural evidence. All panel members were adequately trained, prepared to perform the standard-setting judgments, and found the process easy to follow. The panelists were able to apply their professional experience in making judgments, and most reported being highly satisfied with the recommended proficiency levels. Procedural validity is a crucial criterion for evaluating the quality of standard-setting (Cizek & Bunch, 2007; Kane, 2001; Tannenbaum & Katz, 2013).

The study results showed that the G-TELP Speaking and Writing tests can be used to distinguish between the six proficiency levels of the CEFR scale. The panel members determined that they could distinguish levels from the most basic (A1) to the most advanced (C2) on the CEFR scale. The tests appear to measure a wide range of abilities by including items of varying difficulty. The comparison tables based on the CEFR scale, derived from the post-workshop analysis, showed comparability with other English proficiency tests. This also confirmed that there is little difference compared to previous G-TELP CEFR comparison tables. Furthermore,

the study successfully derived additional levels, such as B1+ and B2+, demonstrating the advanced nature of the CEFR workshop results compared to previous outcomes.

This study assessed the content and methods of the G-TELP in an objective and consistent manner, linking it with the latest language education standards, the CEFR. By adopting the CEFR, it was found that G-TELP Speaking and Writing tests emphasize sociolinguistic knowledge and language strategies, rather than focusing solely on grammar and vocabulary. Based on the level-setting schedule and processes established in this study, it is expected that the G-TELP Level 3, G-TELP Junior, and other tests can also be aligned with the CEFR. Based on these research results, future studies could increase the diversity of panel participants and compare the difficulty levels of different items in the same test to study level-setting with the CEFR. Additionally, studies on the equivalency of the G-TELP Speaking test with other English-speaking tests can be conducted, and the results applied to CEFR-related level-setting to better understand the significance of the scores.

VI. References

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2006). Analysing tests of reading and listening in relation to the common European framework of reference: The experience of the Dutch CEFR construct project. *Language Assessment Quarterly: An International Journal*, *3*(1), 3–30.

Bejar, I. I., Braun, H. I., & Tannenbaum, R. J. (2007). A prospective, progressive, and predictive approach to standard setting. In R. W. Lissitz (Ed.), *Assessing and modeling cognitive development in school: Intellectual growth and standard setting* (pp. 1–30). Maple Grove, MN: Journal of Applied Metrics Press.

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

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education*, *17*, 59–88.

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

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: SAGE.

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. .

Geisinger, K. F., & McCormick, C. M. (2010). Adopting cut scores: Post-standard-setting panel considerations for decision makers. *Educational Measurement: Issues and Practice, 29*, 38–44.

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

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

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 53–88). Mahwah, NJ: Lawrence Erlbaum.

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 with reference to 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.

Milanovic, M. (2009). Cambridge ESOL and the CEFR. *Research Notes, 37*, 2–5.

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

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

Tannenbaum, R. J., & Katz, I. R. (2013). Standard setting. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology: Vol 3. Testing and assessment in school psychology and education* (pp. 455–477). Washington, DC: American Psychological Association.

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.

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing, 22*, 281–300.

Zieky, M. J., Perie, M, & Livingston, S. A. (2008). *Cutscores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

APPENDIX A: Example Round Table Schedule

AGENDA: Mapping G-TELP Test onto the Common European Framework

Day 1: G-TELP Speaking (GST) Section

Day 2: G-TELP Speaking (GST) Section

Day 3: G-TELP Writing (GWT) Section

Day 4: G-TELP Writing (GWT) Section

- 9:00 – 9:45: Table Group: Define groups for each section
- 9:45 – 10:45: Review charts and write explanations for each level sample
- 10:45 – 11:00: Overview of the method for setting response criteria
- 11:00 – 11:15: Break
- 11:15 – 12:00: Training/Practice on standard-setting approaches
- 12:00 – 13:00: Lunch (Data entry)
- 13:00 – 14:00: Round 1 individual decisions on speaking/writing items (starting from lower levels)
- 14:00 – 14:45: Discussion of Round 1 results and score overview
- 14:45 – 15:00: Break
- 15:00 – 15:15: Break (Data entry)
- 15:15 – 15:45: Round 1 individual decisions on speaking/writing items (starting from lower levels)
- 15:45 – 16:00: Discussion of Round 1 results and score overview
- 16:00 – 16:15: Break (Data entry)
- 16:15 – 16:45: Round 1 individual decisions on speaking/writing items

- 16:45 – 16:55: Round 2 individual decisions (overall level for each candidate)
- 16:55 – 17:00: Wrap-up and adjourn

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 : GST Performance Assessment Table by Level

Level		Evaluation Guidelines
1	**Authentic**	The level at which you can speak English fluently at the same level as a native English speaker. You can communicate fluently and logically and have the same pronunciation and accent as native English speakers.
2	**High-Advanced**	Proficiency in English at an advanced level. You can present your opinion on the analyzed information to persuade the other party or perform tasks related to offering a solution to a hypothetical crisis without difficulty. There are occasional mispronunciations, but overall, the expression is smooth and natural. Has a rich vocabulary and can speak with reasonable control of grammatical structures
3	**Advanced**	Able to express one's opinions generally well in almost all situations, although sometimes spontaneously. The flow of speech may be interrupted by starting a sentence incorrectly and rephrasing it. Pronunciation and accent mistakes, as well as intonation and rhythm of the native language, sometimes interfere with conversation.
4	**High-Intermediate**	This is a level at which you can express your opinion in most situations. However, there are solid accents and frequent grammatical mistakes. Often mistakes in intonation make it challenging to convey meaning. Even when we talk, we sometimes say or stop saying unnecessary things.
5	**Intermediate**	Able to express one's opinions in general under normal circumstances but may occasionally have difficulty in unfamiliar situations. Solid accents and frequent grammatical mistakes are present. In addition, it is sometimes difficult to convey the meaning due to apparent errors in stress.
6	**Low-Intermediate**	You can usually communicate your thoughts well in everyday situations, but sometimes it is difficult to respond effectively when dealing with unfamiliar situations. The choice of vocabulary is also generally inappropriate, and there is difficulty in paraphrasing to convey meaning. In addition, differences in stress are evident and often need help to get meaning.
7	**High-Basic**	Even in everyday situations, it is difficult to convey one's opinion, and sometimes it is difficult to answer even in unfamiliar situations. There is a lot of time to think before answering, so the answer is delayed, and the solution needs more. The choice of grammar and vocabulary is also often inadequate.
8	**Basic**	I find it difficult to express my opinion in everyday situations and often find it challenging to respond in unfamiliar situations. Mistakes in frequently used grammatical structures and sentence forms result in little understanding transfer.
9	**Low-Basic**	Difficulty expressing one's opinion in general situations and answering in unfamiliar situations. Responses are always late and need more information. In addition, it is almost impossible to convey understanding due to incorrect grammar and vocabulary.
10	**Beginner-Basic**	It is a level where they have difficulty conveying their thoughts even in familiar everyday situations. Answers are always delayed, and even when they are answered, it is almost impossible to understand. Grammar, vocabulary,

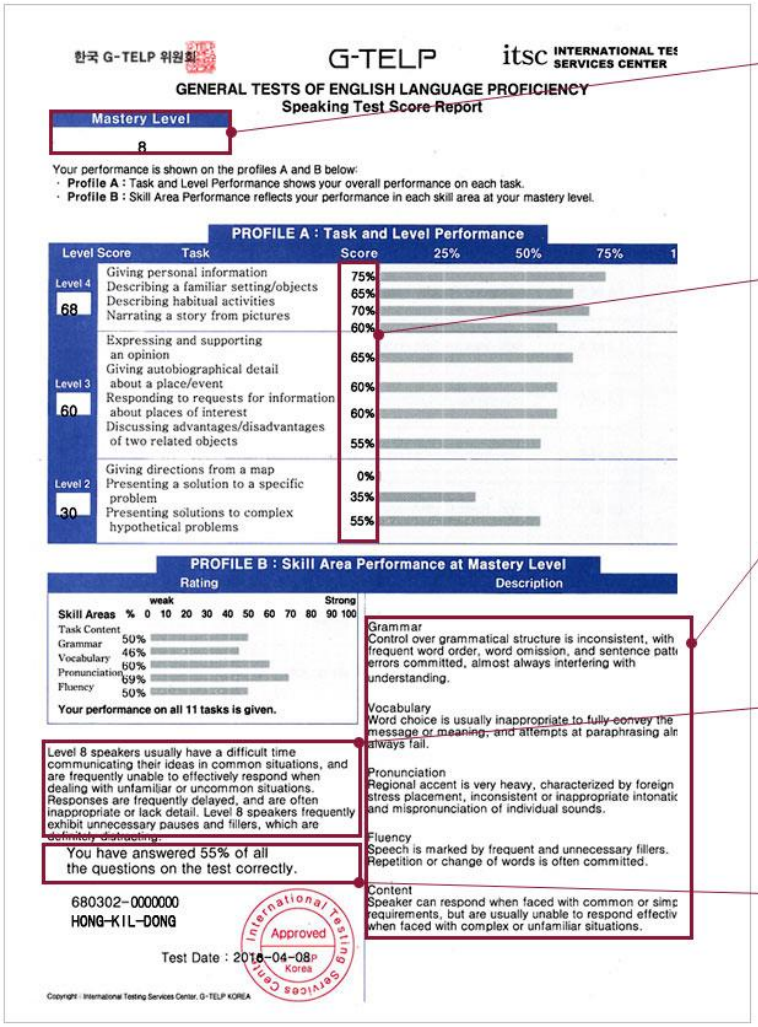
		sentences, stress, everyone makes mistakes all the time, making it impossible to convey meaning.
11	**No mastery**	Unable to hold meaningful conversations, speaking a few words from memory.

APPENDIX E: GST Rating

Rating

<table>
<tr><th rowspan="2">Level</th><th colspan="3">Scores by task</th></tr>
<tr><th>L4
Task1~4</th><th>L3
Task 5~8</th><th>L2
Task 9~11</th></tr>
<tr><td>1</td><td colspan="3">If the average is above 90%</td></tr>
<tr><td>2</td><td colspan="3">If the average is above 75%</td></tr>
<tr><td>3</td><td colspan="2">If the average is above 75%</td><td>If the average is below 74%</td></tr>
<tr><td>4</td><td rowspan="3">If the average is above 75%</td><td>If the average is between 68-74%</td><td rowspan="3">If the average is below 74%</td></tr>
<tr><td>5</td><td>If the average is between 63-67%</td></tr>
<tr><td>6</td><td>If the average is below 62%</td></tr>
<tr><td>7</td><td>If the average is between 70-74%</td><td rowspan="4">-</td><td rowspan="4">-</td></tr>
<tr><td>8</td><td>If the average is between 67-69%</td></tr>
<tr><td>9</td><td>If the average is between 63-66%</td></tr>
<tr><td>10</td><td>If the average is between 55-62%</td></tr>
<tr><td>11</td><td>If the average is below 54%</td><td>-</td><td>-</td></tr>
</table>

APPENDIX F: GST Transcript Sample



한국 G-TELP 위원회

G-TELP

itsc INTERNATIONAL TES SERVICES CENTER

GENERAL TESTS OF ENGLISH LANGUAGE PROFICIENCY
Speaking Test Score Report

Mastery Level
8

Your performance is shown on the profiles A and B below:
- **Profile A :** Task and Level Performance shows your overall performance on each task.
- **Profile B :** Skill Area Performance reflects your performance in each skill area at your mastery level.

PROFILE A : Task and Level Performance

Level Score	Task	Score	25%	50%	75%
Level 4 68	Giving personal information	75%			
	Describing a familiar setting/objects	65%			
	Describing habitual activities	70%			
	Narrating a story from pictures	60%			
Level 3 60	Expressing and supporting an opinion	65%			
	Giving autobiographical detail about a place/event	60%			
	Responding to requests for information about places of interest	60%			
	Discussing advantages/disadvantages of two related objects	55%			
Level 2 30	Giving directions from a map	0%			
	Presenting a solution to a specific problem	35%			
	Presenting solutions to complex hypothetical problems	55%			

PROFILE B : Skill Area Performance at Mastery Level

Rating			Description
Skill Areas	weak % 0 10 20 30 40 50 60 70 80 90 100	Strong	
Task Content			Grammar Control over grammatical structure is inconsistent, with frequent word order, word omission, and sentence patt errors committed, almost always interfering with understanding.
Grammar	50%		
Vocabulary	46%		Vocabulary Word choice is usually inappropriate to fully convey the message or meaning, and attempts at paraphrasing alr always fail.
Pronunciation	50%		
Fluency	69%		Pronunciation Regional accent is very heavy, characterized by foreign stress placement, inconsistent or inappropriate intonatic and mispronunciation of individual sounds.
	50%		Fluency Speech is marked by frequent and unnecessary fillers. Repetition or change of words is often committed.
Your performance on all 11 tasks is given.			Content Speaker can respond when faced with common or simp requirements, but are usually unable to respond effectiv when faced with complex or unfamiliar situations.

Level 8 speakers usually have a difficult time communicating their ideas in common situations, and are frequently unable to effectively respond when dealing with unfamiliar or uncommon situations. Responses are frequently delayed, and are often inappropriate or lack detail. Level 8 speakers frequently exhibit unnecessary pauses and fillers, which are definitely distracting.

You have answered 55% of all the questions on the test correctly.

680302-0000000
HONG-KIL-DONG

International Testing Services Center Approved Korea

Test Date : 2016-04-08

Copyright - International Testing Services Center, G-TELP KOREA

Mastery Level

수험자의 영어 말하기능력 등급을 보여줍니다

Score

세분화된 과제의 점수(task score)각 등급의 세분화된 과제를 어느 정도로 잘 수행하는지를 백분율로 보여줍니 다.

Description

능력묘사표(Description) :
각 기능분야에서의 응시자의 능력을 상세히 설명합니다.

PROFILE B Explanation

응시자의 종합적인 영어말하기 능력수준을 상세히 묘사합니다.

Percent

구술능력을 100점 만점의 평균점수로 표기하여 비교, 활용이 편리합니다.

APPENDIX G: GWT Description

GWT Description

Proficiency Level		Level Description
Level 1	**Authentic**	Test takers at this level can demonstrate their self-confidence in all unfamiliar and familiar situations. They can express their feelings well, use a broad range of appropriate vocabulary words, provide accurate explanations, and express appropriate idioms. They can show consistent and precise grammatical structures, sentence patterns, and word order. Open ideas are logical, sequential, and well-organized so that the message they convey is persuasive.
Level 2	**High-Advanced**	Test takers at this level can effectively express their opinions in almost any situation. There are rarely visible errors, but the grammatical structure and sentences need help understanding their meaning. Nevertheless, they use a broad range of appropriate vocabulary words and explain consistently and effectively. In addition, test takers of this level lay out logical ideas. The writing is generally coherent and persuasive.
Level 3	**Advanced**	Test takers at this level can effectively express their opinions in almost any situation. The test taker's writing is mostly appropriate to the problem. Grammar has little effect on conveying meaning. Essay errors and sentence patterns occasionally appear, but the test taker's writing can usually be easily understood. In general, the test taker selects appropriate words and uses cul-de-sacs when necessary to navigate vocabulary deficiencies. The test taker's writing is generally coherent and persuasive.
Level 4	**High-Intermediate**	Test takers at this level can express their opinions in most situations. Test takers' works are usually appropriate for the problem. Grammatical errors and sentences that occasionally affect the meaning appear sometimes but are generally easy to understand. Usually, test takers choose the right words; when facing vocabulary deficiencies, they find other ways to express their thoughts. There are signs of trying to solve the story logically, and it is generally well-organized. The writing of the examinee is usually well-coordinated but could be more persuasive.
Level 5	**Intermediate**	Test takers at this level express their opinions well on familiar topics, but they need help with writing in unfamiliar subjects. Writing often needs to be on a familiar subject matter. Grammatical errors and sentence patterns that affect meaning sometimes appear but can be primarily understood. Generally, the test taker uses appropriate word choice, but specifically struggles with effective paraphrasing. Ideas are somewhat logical. There are signs of trying to solve ideas logically, and works are generally well-organized. The test taker's writing could be more coherent and more persuasive

Level 6	**Low-Intermediate**	Test takers at this level generally express their opinions well on familiar topics. Still, sometimes it is impossible to communicate effectively in situations that need a detailed explanation. Irrelevant content may be displayed. In general, grammatical structures and sentence patterns are appropriate, but there is evidence of wrong word selection and insufficient amplification to convey the content. Descriptions are presented. the writing could be more logical and it does not clearly express what it wants to develop.
Level 7	**High-Basic**	Test takers at this level usually need help expressing their opinions on familiar topics. Writing effectively in unfamiliar situations is usually impossible. A lack of detailed explanations or relevant content is shown. Grammar significantly affects meaning. Essay structures and sentence patterns often appear. Poor word choice is evident. There is a notably limited range of vocabulary. There is almost no logical development and a lack of compositional effectiveness. Development of the test taker’s ideas can be more precise and more coherent.
Level 8	**Basic**	Explanations are insufficiently detailed, and irrelevant content is often shown. Grammatical errors affecting meaning and sentence patterns are almost always visible. Inappropriate and confusing word choice is usually visible, and it isn't easy to interpret. Unordered ideas. The text is not arranged clearly. The test taker's writing is almost always unclear and contradictory.
Level 9	**Low-Basic**	It is usually impossible to understand meaning. Wrong and confusing word choice is always visible. It is challenging to understand what message the examinee is trying to deliver.
Level 10	**Beginner-Basic**	There is a lack of basic explanation and erratic and inappropriate content development. It can be challenging to understand what test takers are trying to achieve. Grammatical structures and sentence patterns are always unclear. In most cases, word choice is inappropriate. It isn't easy to understand what message the examinee is trying to convey.
Level 11	**No mastery**	Test takers at this level can only express a list of known words or phrases. Therefore, it is impossible to write a composition that conveys the relevant content.

APPENDIX I : GST Transcript Sample

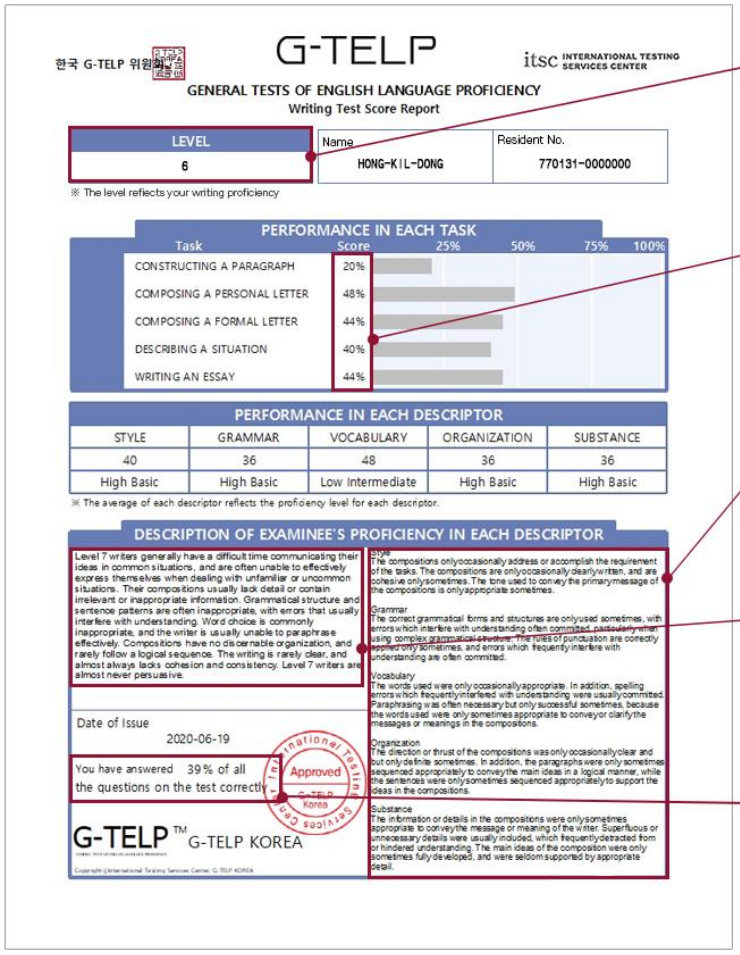
Score Explanation Table

Raw Score	Quality of Sample	% Equivalent
0	no or limited sample	0
1	poor	20
2	fair	40
3	above average	60
4	good	80
5	excellent	100

Grade Conversion Table

Writing Proficiency Level	% Range
1	95 - 100
2	85 - 94
3	75 - 84
4	65 - 74
5	55 - 64
6	45 - 54
7	35 - 44
8	25 - 34
9	15 - 24
10	5 - 14
11	0 - 4

APPENDIX I : GST Transcript Sample



한국 G-TELP 위원회

G-TELP

itsc INTERNATIONAL TESTING SERVICES CENTER

GENERAL TESTS OF ENGLISH LANGUAGE PROFICIENCY

Writing Test Score Report

LEVEL	Name	Resident No.
6	HONG-KIL-DONG	770131-0000000

※ The level reflects your writing proficiency

PERFORMANCE IN EACH TASK

Task	Score	25%	50%	75%	100%
CONSTRUCTING A PARAGRAPH	20%				
COMPOSING A PERSONAL LETTER	48%				
COMPOSING A FORMAL LETTER	44%				
DESCRIBING A SITUATION	40%				
WRITING AN ESSAY	44%				

PERFORMANCE IN EACH DESCRIPTOR

STYLE	GRAMMAR	VOCABULARY	ORGANIZATION	SUBSTANCE
40	36	48	36	36
High Basic	High Basic	Low Intermediate	High Basic	High Basic

※ The average of each descriptor reflects the proficiency level for each descriptor.

DESCRIPTION OF EXAMINEE'S PROFICIENCY IN EACH DESCRIPTOR

Level 7 writers generally have a difficult time communicating their ideas in common situations, and are often unable to effectively express themselves when dealing with unfamiliar or uncommon situations. Their compositions usually lack detail or contain irrelevant or inappropriate information. Grammatical structure and sentence patterns are often inappropriate, with errors that usually interfere with understanding. Word choice is commonly inappropriate, and the writer is usually unable to paraphrase effectively. Compositions have no discernable organization, and rarely follow a logical sequence. The writing is rarely clear, and almost always lacks cohesion and consistency. Level 7 writers are almost never persuasive.

Style
The compositions only occasionally address or accomplish the requirement of the tasks. The compositions are only occasionally clearly written, and are cohesive only sometimes. The tone used to convey the primary message of the compositions is only appropriate sometimes.

Grammar
The correct grammatical forms and structures are only used sometimes, with errors which interfere with understanding often committed, particularly when using complex grammatical structure. The rules of punctuation are correctly applied only sometimes, and errors which frequently interfere with understanding are often committed.

Vocabulary
The words used were only occasionally appropriate. In addition, spelling errors which frequently interfered with understanding were usually committed. Paraphrasing was often necessary but only successful sometimes, because the words used were only sometimes appropriate to convey or clarify the messages or meanings in the compositions.

Organization
The direction or thrust of the compositions was only occasionally clear and but only definite sometimes. In addition, the paragraphs were only sometimes sequenced appropriately to convey the main ideas in a logical manner, while the sentences were only sometimes sequenced appropriately to support the ideas in the compositions.

Substance
The information or details in the compositions were only sometimes appropriate to convey the message or meaning of the writer. Superfluous or unnecessary details were usually included, which frequently detracted from or hindered understanding. The main ideas of the composition were only sometimes fully developed, and were seldom supported by appropriate detail.

Date of Issue
2020-06-19

Approved G-TELP Korea

You have answered 39% of all the questions on the test correctly.

G-TELP™ G-TELP KOREA

Mastery Level

수험자의 영어 작문능력 등급을 보여줍니다.

Score

세분화된 과제의 점수(task score) 각 등급의 세분화된 과제를 어느 정도로 잘 수행하는지를 백분율로 보여줍니다.

Description

능력묘사표 (Descriprion) :
각 기능분야에서의 응시자의 능력을 상세히 설명합니다.

PROFILE B Explanation

응시자의 종합적인 영어쓰기 능력수준을 상세히 묘사합니다.

Percent

쓰기능력을 100점 만점의 평균점수로 표기하여 비교, 활용이 편리합니다.

APPENDIX I: Level-Setting Workshop Panel

Internal Members

Researcher Minjeong Kim

Researcher Hajun Yoo

Manager Sunhee Jeong

Candice Bayley

Corey Steiner

Rob Walsh

Toby Charles William

External members

Ali Raddaoui

Kymberly Talor

Mike Dong