

# Developing a Unified Scoring and CEFR Mapping System for the Integrated GTELP Assessment

**Minjung Kim**

G-TELP Korea

# Table of Contents

1 Background and Research Problem

2 Purpose of the Study

3 Proposed Unified Scoring Framework

4 Challenges of Score Integration

5 CEFR Alignment Framework

6 Planned Statistical Analyses

7 Validity Framework

8 Expected Contributions and Conclusion

# Background

*Language assessment is moving toward integrated communicative competence — stakeholders increasingly want a single, internationally comparable proficiency indicator.*

## GTELP currently reports separately

- Level 2
- Speaking
- Writing

## Separate reporting limits usability in

- Admissions
- Recruitment
- Placement
- International benchmarking

## Key psychometric challenges

Construct representation

Score weighting

Scale compatibility

Validity

Ordinal + continuous integration

# Purpose of the Study

**Main Purpose** Develop a unified scoring system and a CEFR alignment framework for the integrated GTELP assessment.

## Specific Objectives

1

Create a 100-point composite score

2

Develop an appropriate weighting structure

3

Align scores with CEFR levels

4

Improve score reporting transparency

5

Provide psychometric validity evidence

# Proposed Unified Scoring Framework

## Preliminary Unified Scoring Structure

$$\text{Unified Score} = (\text{L2} \times 0.40) + (\text{Speaking} \times 0.30) + (\text{Writing} \times 0.30)$$

**40%**

GTELP Level 2

**30%**

GTELP Speaking

**30%**

GTELP Writing

## Important Considerations

- The weighting structure will be empirically evaluated
- All components will be aligned to a common continuous metric before aggregation
- The framework balances construct representation, reliability, and communicative competence

# Challenges of Score Integration

*Heterogeneous Score Integration — GTLP combines continuous scores with ordinal proficiency levels.*

1

## Preliminary Approach

- Midpoint conversion

2

## Potential Concerns

- Equal-interval assumption
- Reduced score precision
- Boundary misclassification

3

## Psychometric Refinement

- Rasch scaling
- Many-Facet Rasch Measurement (MFRM)
- Empirically derived continuous scale

*Important Principle — midpoint conversion is treated only as a preliminary operational approximation.*

# CEFR Alignment Framework

## Multi-Source Evidence Approach

- Descriptive analyses
- Distributional analyses
- Standard-setting procedures
- External benchmark comparisons
- Expert judgment and validation

## Standard-Setting Methods

- Modified Angoff
- Bookmark procedures

## Important Principle

### CEFR alignment is based on:

- Standard-setting + empirical validation

### NOT:

- Arbitrary score division

## Interpretive Goal

*Interpretive referencing rather than direct score equivalence.*

# Planned Statistical Analyses

## Descriptive Statistics

- Mean
- Standard deviation
- Distributional characteristics

## Reliability Analysis

- Cronbach's alpha
- Inter-rater reliability
- Classification consistency

## Correlation Analysis

- GTLP ↔ TOEFL / IELTS relationships

## Factor Analysis

- EFA, CFA, bifactor models
- Construct representation

## IRT / Rasch Analysis

- Common metric scaling
- Score transformation
- Item functioning



# Validity Framework

## Unified Validity Argument

### Content Validity

Broad coverage of language skills

### Construct Validity

Related but distinct language domains

### Criterion-Related Validity

Relationship with external tests

### Consequential Validity

Improved interpretability and usability

**Important Principle: the unified score should complement sub-scores — not replace them.**

*Key Message — integrated score interpretation requires psychometric coherence rather than simple score aggregation.*

# Contributions and Conclusion

## Practical Contributions

- Improved score reporting
- Greater transparency
- Enhanced usability

## Methodological Contributions

- Psychometrically grounded scoring
- CEFR alignment framework
- Heterogeneous score integration

## Future Implications

- Foundation for further validation
- Future linking & equating studies
- Expanded international usability

*Goal: a transparent, interpretable, internationally meaningful GTELP reporting system.*

# Current GTELP Structure

| Domain    | Item Range |
|-----------|------------|
| Grammar   | 1–26       |
| Listening | 27–52      |
| Reading   | 53–80      |

80

total items across the  
GTLP Level 2 test

# Current Data Structure

## Available Datasets

Separate Level 2 dataset

Separate Speaking dataset

Separate Writing dataset

*Important Limitation — no matched examinee data across all components, so individual-level composite scores cannot yet be directly validated.*

## Study Positioning

- Framework-development study
- Simulation-based evaluation
- Reporting-system proposal

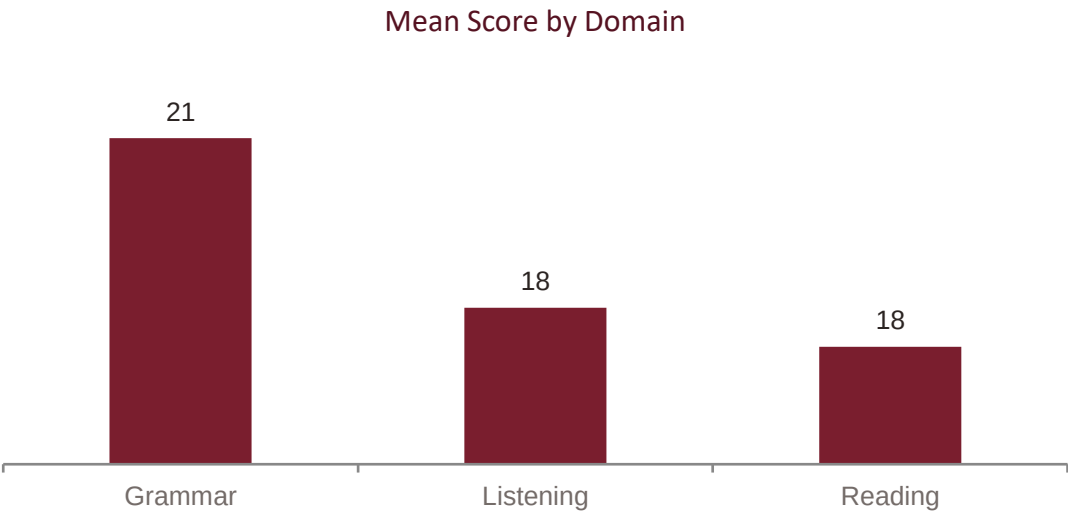
# Empirical Characteristics of Level 2

## Level 2 Descriptive Statistics

| Domain       | Mean Score |
|--------------|------------|
| Grammar      | 21.0       |
| Listening    | 18.4       |
| Reading      | 17.8       |
| Overall Mean | 53.5 / 100 |

## Observed Features

- Broad score distribution
- Balanced receptive-domain representation
- Strong variability across examinees

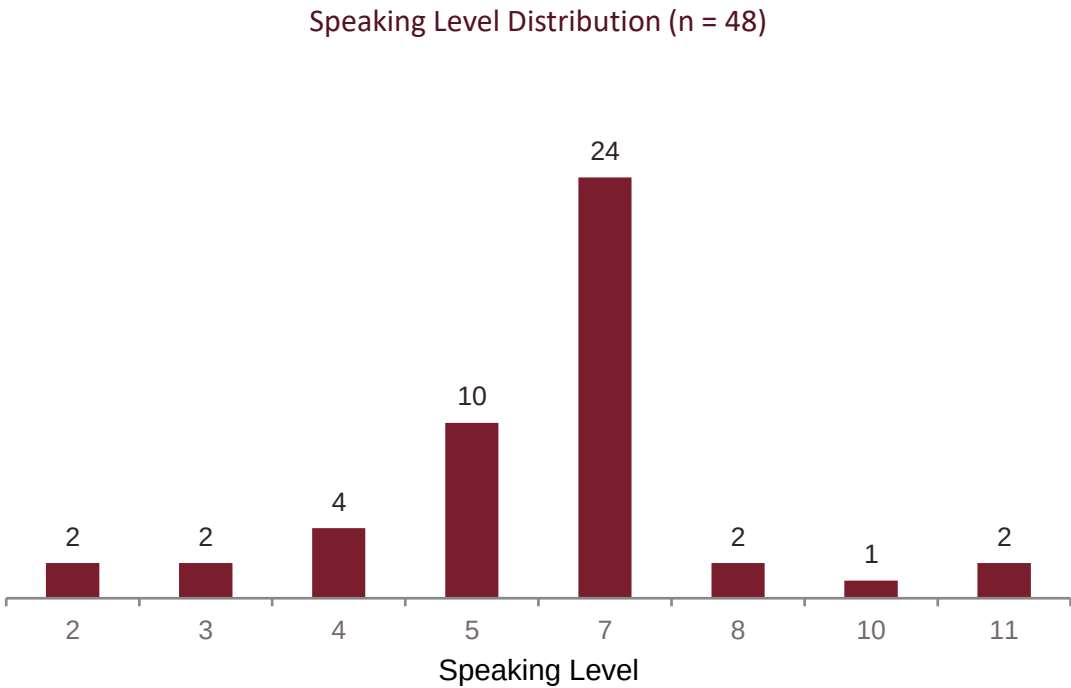


# Speaking Score Variability

| Variable            | Value |
|---------------------|-------|
| Sample Size         | 48    |
| Mean Speaking Level | 4.2   |
| SD                  | 1.1   |

*Key Observation — substantial variability existed even within the same speaking level.*

*Implication — simple midpoint conversion may oversimplify productive-skill performance.*



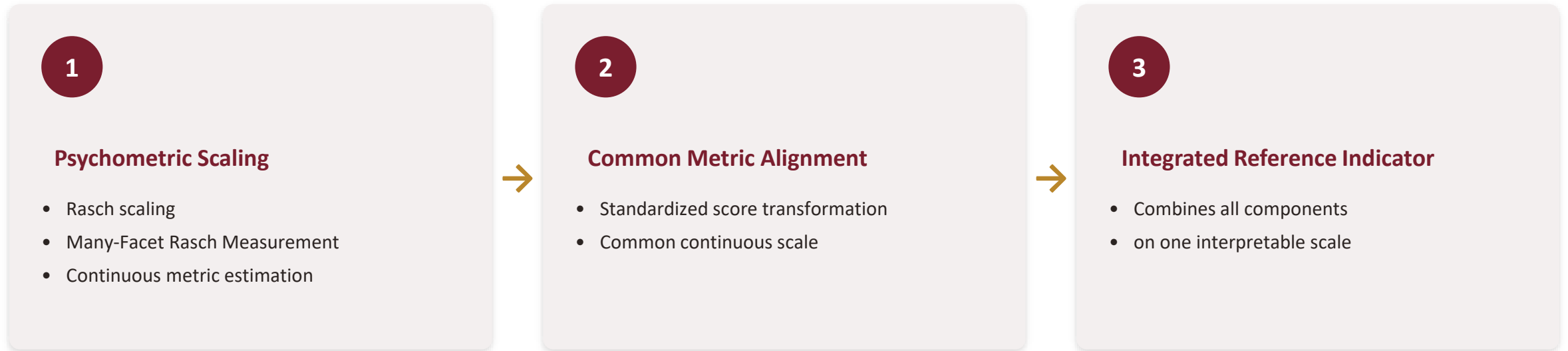
# Example Reporting Structure

| Component | Score   | CEFR |
|-----------|---------|------|
| Level 2   | 82      | B2   |
| Speaking  | Level 5 | B2   |
| Writing   | Level 4 | B1+  |

**Integrated Reference Indicator**

- Supplementary interpretive indicator
- Intended to complement sub-scores
- Supports overall proficiency interpretation

# Revised Integrated Scoring Model



$$\text{Unified Score} = (0.40 \times \text{L2}) + (0.30 \times \text{Speaking}) + (0.30 \times \text{Writing})$$

*Integrated Reference Indicator — a supplementary interpretive indicator that complements sub-scores and supports overall proficiency interpretation.*



# Key Methodological Issues

1

## Heterogeneous Score Integration

- Continuous vs. ordinal scales

2

## Midpoint Conversion Limitations

- Equal-interval assumption
- Reduced score precision
- Potential misclassification

3

## Reliability Differences

- Objective vs. rater-mediated scoring

# Monte Carlo Simulation Design

## Simulation Design

| Scenario | Assumed Correlation |
|----------|---------------------|
| Low      | .30                 |
| Moderate | .50                 |
| High     | .70                 |

### Monte Carlo Simulation Framework

#### Inputs

- Actual Level 2, Speaking, and Writing score distributions

#### Correlation Scenarios

- Low, Moderate, High (see table)

#### Purpose

- Examine the stability and sensitivity of the proposed scoring model

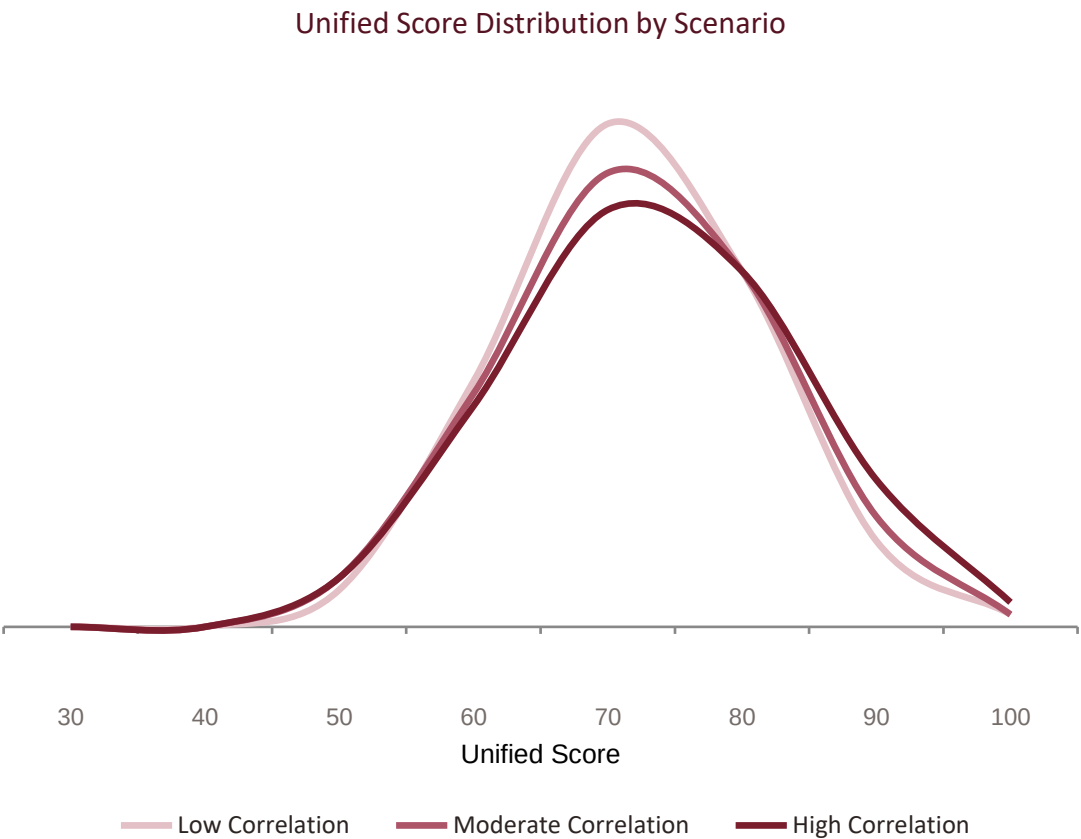
# Simulation Results

## Example Simulation Outcomes

| Scenario | Mean | SD   | % B2+ |
|----------|------|------|-------|
| Low      | 71.8 | 9.6  | 40.8% |
| Moderate | 72.4 | 10.5 | 42.3% |
| High     | 73.1 | 11.3 | 43.7% |

## Key Findings

- Stable score distributions across scenarios
- No excessive dominance by a single component
- Reasonably consistent CEFR classification patterns



# Planned Future Validation

## Future Validation Requirements

- Matched examinee data
- Integrated score validation
- Composite reliability estimation
- Classification consistency

## Planned Analyses

- CFA and bifactor models
- Rasch and MFRM analyses
- Weighting sensitivity analyses
- External validity studies

# Expected Contributions

## Practical Contributions

- Integrated GTELP reporting framework
- Improved interpretability
- Enhanced CEFR usability
- Transparent reporting structure

## Methodological Contributions

- Heterogeneous score integration framework
- Simulation-based evaluation approach
- Psychometric foundation for future validation

# Conclusion

The goal is not simply *score aggregation*  
but rather ***meaningful and defensible score interpretation.***

- The proposed framework appears feasible as a preliminary integrated reporting model
- Simulation results support the stability of the proposed approach
- Full operational validation will require matched examinee data

# Thank You

Minjung Kim · G-TELP Korea · ITSC Research Series ITSC-2025-01