

---

# **New Directions East Asia 2024      Bangkok, Thailand • 21-23 November, 2024**

---

**Submission 140**

## **Evaluating Scoring Validity: Insights from the BESTEP English Proficiency Test in Taiwan**

Lydia H. Y. Lee , Rachel Y. F. Wu , Cecilia H. Y. Liao

*The Language Training & Testing Center*

Scoring validity refers to the degree to which test scores are free from measurement errors and accurately reflect the abilities being measured (Weir, 2005). Ensuring the scoring validity of a language assessment, especially for high-stakes purposes with significant implications for EAP (English for Academic Purposes) teaching and learning, is essential to ensure the reliability and consistency of test scores, thereby providing robust evidence for their appropriate use and interpretation. This study investigates the scoring validity of BESTEP, a test specifically developed for Taiwanese college students and launched in 2023. Aligned with higher education curricula and linked to the Common European Framework of Reference for Languages (CEFR) (Fan et al., 2024), BESTEP assesses college students' general academic English communicative skills, particularly for those participating in English-medium instruction (EMI) courses.

Data from over 5,000 college students' performances on the BESTEP speaking and writing tests were analyzed. The many-facet Rasch model (MFRM) was utilized to evaluate rater effects and test form difficulties, revealing consistent rater severity and parallel test form difficulty. Additionally, analysis of covariance (ANCOVA) was employed to compare the performances of students from different academic disciplines (i.e., liberal arts, science/engineering, and biology/agriculture/medicine). No significant differences were found in their

scores, suggesting that BESTEP does not favor students from any specific majors.

These findings support that the test scores of BESTEP speaking and writing sections serve as reliable and accurate measures of college students' academic English proficiency, providing higher education institutions with a tool to evaluate students' prerequisite skills before and after taking EMI courses. Although qualitative approaches are also necessary to fully establish the scoring validity of a new test such as BESTEP, the findings of this quantitative investigation contribute to establishing fundamental evidence for scoring validity and enhancing test transparency for its stakeholders.