Executive Summary

Mapping BESTEP Listening and Reading Test Scores to CEFR Levels Using Test-Centered and Statistical Methods

Sun-Young Shin & Sijia Huang
Indiana University

This project examined the alignment of the Listening and Reading sections of the BEST Test of English Proficiency (BESTEP) with the Common European Framework of Reference for Languages (CEFR) to strengthen the interpretability and international comparability of BESTEP scores. The study was conducted as part of the LTTC Teaching and Research Grants program and aimed to provide empirical evidence to support the continued use and refinement of CEFR-linked score interpretations for BESTEP within Taiwan's bilingual education and English-medium instruction (EMI) initiatives. To ensure both methodological rigor and practical relevance, the study combined expert judgment with complementary statistical analyses.

A four-round Bookmark standard-setting procedure was implemented with 18 panelists from two independent panels located at LTTC in Taiwan and Indiana University in the United States. Panelists reviewed BESTEP Listening and Reading items organized in an Ordered Item Booklet constructed from operational test data from 1,674 test takers using item response theory calibration. The results demonstrated strong consistency in panelists' judgments within and across the two sites, indicating that the proficiency thresholds were interpreted similarly across panels with different professional and institutional backgrounds. This high level of agreement provides strong support for the stability and credibility of the standard-setting process and confirms that BESTEP Listening and Reading scores reflect a coherent progression of proficiency consistent with CEFR descriptors.

The final reconciled cut scores indicate that BESTEP Listening and Reading effectively classify test takers across CEFR levels from A1 through B2+. The score ranges above the B2 threshold represent higher performance within the B2 band. These results demonstrate that the test provides meaningful differentiation across the proficiency range most relevant to university students participating in EMI programs. Overall, the findings confirm that BESTEP functions as a reliable tool for assessing academic English readiness within the intended proficiency range of the BEST program.

To complement the expert judgment process, hierarchical cluster analysis and latent class analysis were conducted using the same dataset to examine whether statistically derived proficiency groupings corresponded with the panel-based cut scores. The results showed generally comparable patterns of proficiency progression, particularly within the intermediate proficiency bands where most BESTEP test takers are located. The convergence between expert judgment and statistical analyses provides additional empirical support for the validity of the CEFR alignment and strengthens the overall evidence base supporting the current proficiency framework.

Comparisons with the currently operational BESTEP thresholds indicate that the existing CEFR framework used by LTTC is broadly consistent with the patterns observed in the present study. While some differences in cut score locations were observed, the hierarchical ordering of CEFR levels remained stable across all methods and analyses. These results provide reassurance that the current reporting structure reflects the overall proficiency continuum measured by BESTEP and supports transparent interpretation of test scores across educational contexts.

Based on these findings, the study recommends continuing the current CEFR-aligned reporting approach while considering several future enhancements that could further strengthen the BESTEP framework. Periodic replication of CEFR alignment studies using additional operational test administrations would help confirm the stability of cut scores over time.
Expanding the item pool across proficiency levels, particularly at higher levels, may also support even finer distinctions in proficiency should LTTC wish to expand reporting in the future. In addition, by integrating expert judgment with statistical analyses, BESTEP will maintain multiple validation approaches, ensuring the program consistently meets international best practices in language assessment.

Overall, the study provides strong empirical evidence supporting the CEFR alignment of BESTEP Listening and Reading scores and reinforces the role of BESTEP as a credible and internationally interpretable measure of English proficiency for university students in Taiwan. The findings also offer practical guidance that will assist LTTC in maintaining and further strengthening the transparency, validity, and policy relevance of BESTEP score interpretations in the years ahead.

中文摘要

◆ 研究團隊與研究目的

本研究由美國印第安納大學布盧明頓分校（Indiana University Bloomington）Sun-Young Shin教授與Sijia Huang教授主持，依照*Aligning Language Education with the CEFR: A Handbook*（British Council et al., 2009）以及 *Relating Language Examinations to the CEFR: A Manual*（Council of Europe, 2009）建議的程序—包含熟悉CEFR分級（familiarization）、測驗內容分析（specification，即審視測驗品質與內容和CEFR級數的關聯）、標準設定（standardization，即判斷試題對應的CEFR級數）、與實證研究（empirical validation）等四階段—由布盧明頓及臺北兩地組成的專家小組（twin panel），判斷「培力英檢」聽力及閱讀能力測驗與CEFR級數之對應關係。研究結果提供「培力英檢」聽力與閱讀能力測驗更多效度證據，且提供增進測驗品質的建議。

◆ 研究問題

1. 對應「培力英檢」聽力與閱讀能力測驗與 CEFR 級數。
2. 比較專家小組中熟悉「培力英檢」者與不熟悉「培力英檢」者判斷結果
3. 探討專家小組判定結果與統計推導結果之間的一致性。

◆ 研究方法摘要

1. 測驗內容分析由印第安納大學與 LTTC 研究團隊協力進行，檢視「培力英檢」聽力與閱讀測驗各部分題型與試題內容，並根據分析結果判定各部分所對應的 CEFR 級數。
2. 標準設定（standardization）由 18 位具語言學、英語教學或語言評量背景的教師與研究人員所組成的兩個專家小組分別在布盧明頓與臺北兩地進行。布盧明頓組不熟悉「培力英檢」，具語言學、英語教學專長與多元的母語背景，並熟悉 CEFR；臺北組熟悉「培力英檢」與 CEFR，具語言學/英語教學與語言評量專長與。標準設定的程序透過線上會議方式進行，以「書籤標準設定程序（Bookmark standard-setting procedure）」，每位成員依試題內容，根據 CEFR 聽力與閱讀能力說明，判斷 CEFR 級數。

◆ 研究結果摘要

1. 根據測驗內容分析和標準設定的結果，「培力英檢」聽力與閱讀測驗能有效區分 CEFR A1 至 B2+ 各能力等級。
2. 內外兩組專家小組的判斷具高度一致性，表示不同背景的專家小組對能力門檻的判定相當一致。此外，專家小組判定結果與統計推導結果大致相符，為「培力英檢」對應 CEFR 的效標參照效度提供實證。
3. 整體而言，研究結果提升了「培力英檢」聽力與閱讀測驗成績的透明度（transparency）與可詮釋性（interpretability），透過對接國際通用的語言能力分級架構 CEFR 級數的實證研究，證明測驗內容與難度符合國際研究標準。