培力英語能力檢定測驗

BEST Test of English Proficiency

信效度報告

Reliability and Validity Report

110 年 8 月至 114 年 10 月

ITTC® 財團法人 語言訓練測驗中心
THE LANGUAGE TRAINING & TESTING CENTER

摘要

「培力英語能力檢定測驗（BEST Test of English Proficiency）」，簡稱「培力英檢（BESTEP）」，係教育部（Ministry of Education, MOE）補助語言訓練測驗中心（The Language Training and Testing Center, LTTC）研發的全國性測驗，評量國內大專校院學生的英語溝通能力。「培力英檢」於 112 年 9 月正式推出，以研究導向設計，提供具信效度、符合國內高等教育情境與需求的英語測驗，評量的程度涵蓋 CEFR A1 至 C1。

「培力英檢」透過帶動學習、教學與評量間的良性循環，培養大專校院學生具備進入學術、職場環境的英語溝通能力。測驗的成績除了可幫助大專校院學生明確了解自己相對於其他大專校院學生的英語能力高低，規劃自主學習，在畢業前備妥足夠的英語力，提升求職與升學的競爭力；也可以提供老師作為了解學生英語能力之參考，據此調整授課方式與內容，以提升教學成效。教育部與相關決策單位亦可以參考成績規劃提升英語力的配套措施，從而整合資源，擴散效益，系統化帶動大專校院學生發展英語能力。截至 113 年底，報考人次約 40,000 人。

「培力英檢」品質深受各界信賴，成績已獲下列國內外機構採認：

1. 教育部「公費留考」與「對外華語教學能力認證考試」的英語能力合格認定基準。
2. 臺灣多所大專校院評估學生修習英語授課課程（EMI）前的英語能力合格認定基準。
3. 奧地利、法國、德國、義大利、馬來西亞、波蘭、瑞士及土耳其等國家的十二所國外大學的英語能力合格認定基準。

「培力英檢」的效度與信度透過多元的研究接軌國際標準，特別是《歐洲語言共同參考架構》（CEFR）。「培力英檢」的信效度研究與「全民英檢」（GEPT）信效度研究同樣參考 Weir（2005）的「社會認知效度驗證框架」進行，包括認知與語境效度（即構念效度）、計分/評分效度、效標關聯效度、後果效度等，全面、系統化蒐集「培力英檢」的效度與信度資訊。

英國文化協會測評研發中心 Barry O’Sullivan 教授特別對於「培力英檢」的發展與研究成果給予高度的肯定，表示「培力英檢」是專為臺灣高等教育情境設計的高品質評量工具，並與 CEFR 等國際標準接軌。雖然推出僅僅三年，對臺灣大

專校院英語學習已經發揮顯著的正面影響，這樣的成就通常需要更長時間才能達成。LTTC 致力於研發符合臺灣需求的在地化測驗的努力，足以作為其他國家發展在地化語言測驗的典範。

以下摘要「培力英檢」主要的信效度研究：

1. 與 TOEFL iBT 的信度比較：「培力英檢」與 TOEFL iBT 各項測驗信度指數相當，符合國際標準。
2. 與 CEFR 級數對接：聽、說、讀、寫均參考歐洲理事會（Council of Europe, 2009）建議的程序，對接 CEFR 級數。
3. 對學術英語學習的影響：探討「培力英檢」成績與大專校院學生學術英語表現的關聯性，確認測驗能有效追蹤評量學生的英語溝通能力。
4. 評分效度與公平性：分析成績是否受到非英語能力（例如主修、年級、性別、考試場次等）因素影響，造成偏誤，確認成績的公平性。
5. 語料庫研究：建置口說作答語料庫，為未來研發 AI 自動評分模型做準備。

「培力英檢」的成功經驗獲得國際專家的肯定，不僅展現其作為全球認可並符合在地需求的英語能力測驗的價值，也反映其未來發展的潛力。

目次（中文版）

一、前言

本報告彙整了由國內外研究團隊參考 Weir（2005）的測驗效度驗證框架進行的「培力英檢」信效度研究，系統化評估各面向的測驗效度，反映測驗符合在地需求，與國際標準接軌。

英國文化協會測評研發中心 Barry O'Sullivan 教授高度肯定「培力英檢」是一項高品質、符合國際標準的測驗，並強調：

- 雖然開辦僅三年，「培力英檢」對臺灣高等教育中的英語學習已發揮強大影響力。
- LTTC 具備測驗在地化以及與國際接軌的專業能力。
- 「培力英檢」可以作為其他國家發展在地化語言測驗的典範。

二、信度分析

「培力英檢」從研發階段到正式辦理測驗，監控測驗的信度一向是製卷標準作業程序的重要一環，包括內部一致性信度、複本信度、評分者間信度與評分者內部信度以及測驗公平性。

「培力英檢」的信度與 TOEFL iBT 在 109 年公布的《TOEFL® Research Insight Series 期刊》第三卷各項信度指數大致相當，符合國際測驗標準，包括：

- 內部一致性信度與 TOEFL iBT 相當。
- 不同場次測驗的信度維持穩定。
- 口說與寫作測驗的評分者內部信度與評分者間信度一致性相當高。

表一：信度指數[1]

測驗	分數範圍	內部一致性信度	測量標準誤
聽力	0 – 140	0.87 – 0.91	8.88 – 9.47
閱讀	0 – 140	0.86 – 0.92	9.27 – 9.38
測驗	分數範圍	評分者間信度	評分者內部信度
寫作	0 – 360	0.84 – 0.85	0.94 – 0.95
口說	0 – 360	0.85 – 0.87	0.95 – 0.97

三、效度研究

LTTC 與國內外大學及研究機構合作，參考 Weir（2005）的效度驗證框架進行以下多面向的「培力英檢」的效度研究。

1. CEFR 對接研究

為了讓成績使用者與國內外學界進一步了解「培力英檢」的品質與分數的意義，LTTC 與國外研究團隊合作完成了「培力英檢」的說寫測驗對接研究，同時也自行進行聽讀測驗對接研究中：

- 口說測驗與 CEFR 對接研究：由英國貝德福德大學（University of Bedfordshire）與 LTTC 合作於 112 至 113 年期間進行，循歐洲理事會建議的程序，分析口說測驗任務與信效度，以及分數與 CEFR 級數之間的關係。
- 寫作測驗與 CEFR 對接研究：由澳洲墨爾本大學（University of Melbourne）與 LTTC 合作於 112 至 113 年期間進行，同樣循歐洲理事會建議的程序，分析寫作測驗任務與 CEFR 能力指標的相符程度，以及寫作測驗分數與 CEFR 級數的對應關係。
- 聽力及閱讀測驗與 CEFR 對接研究：由美國印第安那大學布盧明頓分校（Indiana University Bloomington）於 114 至 115 年間進行，循歐洲理事會建議的程序，採四階段驗證流程。

[1] TOEFL iBT 信度估計及標準誤

分數	範圍	信度估計	標準誤
閱讀	0–30	0.87	2.34
聽力	0–30	0.87	2.38
口說	0–30	0.86	1.57
寫作	0–30	0.80	2.14

2. 對學術英語學習的影響研究

國立臺灣大學（NTU）與國立臺灣科技大學（NTUST）於 112 至 113 年間進行的一項研究，探討以下研究問題：

- 「培力英檢」分數與大學學術英語課程成績的關聯性
- 學生在課程中的英語能力進步幅度
- 「培力英檢」與學習目標、評量任務及評分標準的關係

研究結果證實，在「培力英檢」中表現優異的學生，在學術英語學習中也展現顯著的進步。

3. 寫作測驗構念效度研究

英國貝德福德大學（University of Bedfordshire）於 112 至 114 年參考 Weir（2005）效度驗證框架進行研究，探討以下研究問題：

- 「培力英檢」寫作測驗任務的語境，與目標語使用的相符程度
- 「培力英檢」寫作測驗任務所引導學生的認知過程
- 學生及 EMI 教師對「培力英檢」的看法

這項研究進一步證明「培力英檢」的構念（construct）效度。

4. 口說測驗語料庫建置

東京外國語大學（Tokyo University of Foreign Studies）與 LTTC 研究團隊合作，於 113 至 115 年期間建置口說語料庫，研究目標包括：

- 分析不同 CEFR 級數與「培力英檢」分數考生口說語料的特徵。
- 作為未來強化評分效度，提供考生更具體的成績回饋，以及發展 AI 自動化評分系統的基礎。

四、國際學術發表

「培力英檢」從研發階段起積極在國內外學術會議發表，與學界交流：

- 111年11月泰國:「亞洲英語語言測驗學術論壇」(Academic Forum on English Language Testing in Asia）
- 112 年 4 月臺灣：「臺灣高教雙語教育論壇」（Bridging Forward）
- 112 年 6 月美國：「國際語言測驗研究年會」（Language Testing Research Colloquium, LTRC）
- 112 年 9 月日本：「亞洲語言測驗協會」(Asian Association for Language Assessment, AALA)
- 112 年 10 月越南：英國文化協會主辦的「英語評量新發展方向」國際研討會（British Council “New Directions in Language Assessment”）
- 113 年 3 月臺灣：「臺灣高教雙語教育論壇」
- 113 年 4 月臺灣：「應用語言學暨語言教學國際研討會」
- 113 年 7 月奧地利：「國際語言測驗研究年會」（Language Testing Research Colloquium, LTRC）
- 113 年 11 月泰國：英國文化協會主辦的「英語評量新發展方向」國際研討會（British Council “New Directions in Language Assessment”）
- 114 年 6 月：泰國「國際語言測驗研究年會」（Language Testing Research Colloquium, LTRC）

五、結論與未來發展

在國內外研究團隊信效度研究結果的支持下，「培力英檢」展現 Weir（2005）驗證框架中的認知與語境效度、計分/評分信效度、效標效度、後果效度均達國際測驗評量界的專業標準。

測驗品質與對英語學習的影響力也受到英國 Barry O’Sullivan 教授的高度肯定，足可作為測驗全球在地化（glocalization）的典範。未來的研究將持續強化以下面向：

- AI 輔助自動評分系統
- 電腦化測驗系統
- 四項測驗完整對接 CEFR
- 縱向研究長期追蹤英語學習進展

「培力英檢」能夠快速融入臺灣高等教育的雙語教育政策，成為全球認可且在地化的優質英語能力測驗，並為未來持續發展奠定穩固基礎。

六、參與計畫人員簡歷

姓名	職稱	最高學歷	執行計畫分工
李欣穎	執行長	美國 State Univ. of New York at Buffalo 英文系博士	計畫主持人
吳若蕙	副執行長兼研發長	英國 Univ. of Surrey Roehampton 語言測驗博士	共同主持人
吳怡芬	測驗編審處 處長	英國 Univ. of Bedfordshire 語言測驗博士	共同主持人
馬冬梅	執行長辦公室 執行長特助	美國 State Univ. of New York at Buffalo 英語教學碩士	計畫聯絡人
施怡芃	綜合測驗處 處長	美國 Univ. of Hawaii at Manoa 英語教學碩士	正式施測試務、閱卷闈場規劃與管理
林君文	教學訓練處 處長	美國 Columbia Univ. 英語與比較文學碩士	評量與學習相關資源規劃與出版
張慶佳	資訊處 處長	國立臺灣大學 法律系學士	行政與試務系統規劃與建置

Executive Summary

The BEST Test of English Proficiency (BESTEP) was developed by the Language Training and Testing Center (LTTC) under the auspices of Taiwan's Ministry of Education (MOE) as a national assessment to evaluate the English proficiency required by students in higher education in Taiwan. Launched in September 2023, the test was designed using a research-driven approach to provide a reliable and contextually relevant measure of English proficiency at CEFR levels A1 to C1.

BESTEP aims to foster a virtuous cycle of learning, teaching, and assessment to enhance students' English communication skills in academic and professional settings. It also helps teachers and institutions tailor their instruction and serves as an indicator of workplace communication ability. By the end of 2024, nearly 40,000 test-takers had taken BESTEP.

BESTEP results are recognized for various academic and professional purposes by the following institutions:

1. The MOE for awarding both the MOE Fellowship for Studying Abroad and the Certification of Proficiency in Teaching Chinese as a Second/Foreign Language.
2. Colleges and universities in Taiwan for evaluating students' readiness for EMI courses.
3. A total of 12 universities in Austria, France, Germany, Italy, Malaysia, Poland, Switzerland, and Turkey.

The validity and reliability of BESTEP have been extensively studied to ensure its alignment with international standards, particularly the Common European Framework of Reference for Languages (CEFR). BESTEP validation work has primarily followed Weir's (2005) socio-cognitive validation framework, covering cognitive, contextual, scoring, criterion-related, and consequential validity. This was also the foundation of the GEPT (General English Proficiency Test) validation research. This methodological approach ensures a rigorous and systematic assessment of BESTEP's validity and reliability.

A recent testimonial from Professor Barry O'Sullivan OBE (see Appendix 2), a distinguished language assessment expert, recognizes BESTEP as "a quality assessment

tool, thoughtfully aligned with both Taiwan's higher education context and international frameworks such as the CEFR. Despite being launched only three years ago, BESTEP has already made a significant positive impact on English learning at the tertiary level—an achievement that often takes much longer to materialize. The LTTC's dedicated research and development efforts in localizing the test to Taiwan's needs serve as an exemplary model for other nations."

Key findings from validation studies include the following:

1. Reliability Comparable to TOEFL iBT: BESTEP has been benchmarked against TOEFL iBT's reliability standards, ensuring consistency in its assessment outcomes.
2. Alignment with CEFR: Speaking, writing, listening, and reading components have been studied to ensure their validity in mapping to CEFR levels.
3. Impact on Academic English Learning: Studies have explored the relationship between BESTEP scores and students' academic English performance, confirming its effectiveness in tracking language development.
4. Scoring Validity and Fairness: Investigations into rating reliability and the impact of test-takers' academic backgrounds show no inherent biases, reinforcing BESTEP's fairness as an assessment tool.
5. Corpus-Based Research and AI Integration: Ongoing projects aim to develop a spoken response corpus to enhance future AI-driven scoring models.

BESTEP's success story, as noted by international experts, highlights its role as a globally recognized, locally optimized English proficiency test that is well-positioned for future growth.

Table of Contents (English Version)

I. Introduction

BESTEP is designed to evaluate English language proficiency within Taiwan's bilingual education framework. This report presents findings from validity and reliability studies conducted by multiple research teams, assessing the test's effectiveness, alignment with international standards, and impact on English learning.

BESTEP's validation work has primarily followed Weir's (2005) test validation framework. This framework provides a systematic methodology for evaluating test validity across multiple dimensions, ensuring that the test meets both local and international assessment standards.

Furthermore, Professor O'Sullivan OBE has endorsed BESTEP as a high-quality, internationally aligned assessment tool. In his recent testimonial, he highlights the following:

- BESTEP's strong impact on English learning in higher education, despite its relatively short history.
- The LTTC's expertise in localizing the test to Taiwan's educational context while maintaining global relevance.
- BESTEP's potential as a model for other countries seeking to develop localized language assessments.

II. Reliability Analysis

BESTEP routinely monitors reliability as part of its standard operating procedures, assessing internal reliability, parallel-form reliability, inter- and intra-rater reliability, and test fairness.

BESTEP's reliability has been benchmarked against TOEFL iBT's 2020 reliability data, demonstrating that its score consistency meets international assessment standards. The study, referencing the *TOEFL® Research Insight Series*, Volume 3, highlights the following:

- BESTEP's internal consistency measures align closely with TOEFL iBT.

- The test's reliability coefficients remain stable across different test administrations.
- Speaking and writing assessments show high inter- and intra-rater reliability, with strong agreement among and within trained human raters.

Table 1. Reliability Estimates[2]

Test	Scale	Cronbach's Alpha	Standard error of measurement
Listening	0 – 140	0.87 – 0.91	8.88 – 9.47
Reading	0 – 140	0.86 – 0.92	9.27 – 9.38
Test	Scale	Inter-rater reliability	Intra-rater reliability
Writing	0 – 360	0.84 – 0.85	0.94 – 0.95
Speaking	0 – 360	0.85 – 0.87	0.95– 0.97

III. Validation Research

BESTEP's validity studies have been conducted in collaboration with domestic and international universities and research institutions. The following major projects have been undertaken using Weir's (2005) validation framework.

1. CEFR Alignment Studies

To ensure alignment with the CEFR, three separate studies have been completed or are ongoing. These are as follows:

- Speaking Test CEFR Mapping – Conducted by the University of Bedfordshire (2023-2024). A panel of experts evaluated the speaking tasks and assessed the relationship between BESTEP speaking scores and CEFR levels.

[2] TOEFL iBT's Reliability Estimates and Standard Error of Measurement

Score	Scale	Reliability Estimate	SEM
Reading	0–30	0.87	2.34
Listening	0–30	0.87	2.38
Speaking	0–30	0.86	1.57
Writing	0–30	0.80	2.14

- Writing Test CEFR Mapping – Conducted by the University of Melbourne (2023-2024). This study evaluated the comparability of writing tasks to CEFR descriptors and the relationships between BESTEP writing scores and CEFR levels.
- Listening and Reading CEFR Mapping – Conducted by Indiana University Bloomington (2025-2026). This study also follows CEFR's four-phase validation process to investigate the relationships between BESTEP listening and reading scores and CEFR levels.

2. Impact Study on Academic English Learning in Higher Education

A study by National Taiwan University (NTU) and National Taiwan University of Science and Technology (NTUST) (2023-2024) investigated the following:

- The relationship between BESTEP scores and university academic English courses.
- The progression of students' English proficiency throughout their coursework.
- The alignment of learning objectives, assessment tasks, and scoring standards.

Findings confirmed that students who performed well on BESTEP also demonstrated measurable English improvement in their academic studies.

3. Validation Study of the Writing Component

Following Weir's (2005) framework, a study by the University of Bedfordshire (2023-2025) examines the following:

- The context validity of the BESTEP writing tasks in relation to the target language use domain.
- The construct of cognitive processes elicited by the BESTEP writing tasks.
- Students' and EMI teachers' perceptions of BESTEP.

This study reinforces that BESTEP is effectively measuring the intended language constructs.

4. Speech Corpus Construction

In collaboration with Tokyo University of Foreign Studies (2024-2026), a spoken response corpus is being developed with the following aims:

- To analyze characteristics of responses at different CEFR levels.
- To enhance the scoring validity, provide test takers with more specific feedback, and support future AI-driven assessments.

IV. Conference Presentations

BESTEP's research has been recognized internationally, with findings presented at conferences such as the following:

- 2022 Academic Forum on English Language Testing in Asia, Thailand
- 2023 Bridging Forward, Taiwan
- 2023 Language Testing Research Colloquium (LTRC), USA
- 2023 British Council New Directions in Language Assessment, Vietnam
- 2024 Taiwan Higher Education Bilingual Education Forum, Taiwan
- 2024 International Conference on Applied Linguistics and Language Teaching, Taiwan
- 2024 Language Testing Research Colloquium (LTRC), Austria
- 2024 British Council New Directions in Language Assessment, Thailand
- 2025 Language Testing Research Colloquium (LTRC), Thailand

These events have validated BESTEP's research efforts within the global language testing community.

V. Conclusion and Future Directions

BESTEP has demonstrated strong validity, reliability, and fairness in assessing English proficiency. By following Weir's validation framework, the research ensures the following:

- The test's construct validity is well-established.
- The contextual factors affecting language performance are well understood.

- Scoring validity and reliability are rigorously examined.

Furthermore, Professor Barry O'Sullivan's endorsement underscores BESTEP's rapid impact, highlighting it as a model for test glocalization. Ongoing research initiatives will further enhance the following:

- AI-assisted automated scoring systems
- Computerized test delivery
- Expanded CEFR alignment studies for all test components
- Longitudinal studies tracking English learning progress

BESTEP's rapid integration into Taiwan's bilingual education strategy at the tertiary level highlights its potential as a globally recognized, locally optimized English proficiency test that is well-positioned for future growth.

VI. References

(For studies marked with an asterisk, an abstract is available in Appendix 1.)

*Chan, S., Nakatsuhara, F., & Jones, J. (2025). *Exploring the context and cognitive validity of BESTEP in an EMI HE context* [Manuscript in preparation]. Language Training and Testing Center.

Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, and assessment (CEFR): A manual.*

Educational Testing Service. (2020). *Reliability and comparability of TOEFL iBT scores* (TOEFL Research Insight Series, Vol. 3, 3rd ed., retrieved from https://www.ets.org/pdfs/toefl/toefl-ibt-insight-s1v3.pdf).

*Fan, J., Chen, I., & Knoch, U. (2024). *Linking the BEST Test of English Proficiency (BESTEP) Writing Test to the Common European Framework of Reference (CEFR)*. Language Training and Testing Center.

*Green, A., & Inoue, C. (2024). *Relating the BESTEP Speaking Test to the Common European Framework of Reference for Languages*. Language Training and Testing Center.

*Lee, L. H. Y., Wu, R. Y. F., & Liao, C. H. Y. (2024). *Evaluating scoring validity: Insights from the BESTEP English Proficiency Test in Taiwan* [Paper presentation]. New Directions East Asia 2024, Bangkok, Thailand.

*Lin, A. C., Wu, R. Y. F., & Wu, J. R. W. (2025). Aligning EAP teaching, learning, and assessment: A case study from Taiwan. International Journal of English for Academic Purposes: Research and Practice, 5(2), 241–281.

*Shin, S & Huang, S. (n.d.). *Mapping BESTEP Listening and Reading Test Scores to CEFR Levels Using Test-Centered and Statistical Methods* [Ongoing project].

*Tono, Y. (n.d.). *BESTEP learner corpora development project*. Language Training and Testing Center [Ongoing project].

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan.

*Wu, J. R. W. (2023). *Assessment in learning systems in Asia—Impeding, driving or improving?* [Panel discussion]. New Directions 2023, Hanoi, Vietnam.

*Wu, J. R. W., Huang, D. H. T., Hung, A. S. T., Lin, A. C. W., Chin, J. S., & Ke, A. S. H. (2024). *A collaborative approach to examining BESTEP's impact on tertiary EAP in Taiwan* [Paper presentation]. 2024 Language Testing Research Colloquium (LTRC), Innsbruck, Austria.

*Wu, J. R. W., Lin, A. C. W., & Chin, J. S. (2024). *Enhancing learning, teaching, and assessment: The interconnection of BESTEP tests and EAP courses in Taiwan* [Paper presentation]. 2024 International Conference on Bilingual Education, Taipei, Taiwan.

*Wu, J. R. W., Wu, R. Y. F. & Lin, A. C. W. (2023). *Co-constructing with stakeholders: The performance descriptors for an English productive skills test* [Paper presentation]. The 44th Language Testing Research Colloquium (LTRC), New York City, NY, United States.

*Wu, J. R. W., Wu, R. Y. F., Lin, A. C. W., & Chin, J. S. (2024). *The BEST Test of English Proficiency (BESTEP): Contextual needs, realization, and future directions* [Panel discussion]. 2024 International Conference on Applied Linguistics & Language Teaching (ALLT), National Taiwan University of Science and Technology.

*Wu, J. R. W. & Wu, R. Y. F., & Liu Anne. (2022). *Developing an EAP test for college students in support of Taiwan's bilingual 2030 policy* [Paper presentation]. Academic Forum on English Language Testing in Asia (AFELTA), Bangkok, Thailand.

Appendix 1. Abstracts of Key Research Studies

Chan, S., Nakatsuhara, F., & Jones, J. (2025). Exploring the context and cognitive validity of BESTEP in an EMI HE context [Manuscript in preparation]. Language Training and Testing Center.

The study employs a mixed-methods approach to evaluate the validity of the BESTEP Writing test (Tasks 2 and 3) in Taiwan's EMI higher education (HE) context. Three key areas are investigated: context validity, cognitive validity, and teachers' perceptions.

Context validity is assessed through a test scrutiny method, analyzing 70 key parameters. The results show that Task 2 focuses on describing information and expressing opinions, while Task 3 emphasizes academic summarization. The BESTEP is tailored to the Taiwanese EMI HE context, ensuring relevance to students' academic needs.

Cognitive validity is explored through a large-scale questionnaire involving 344 test-takers at CEFR B1–C1 levels. Factor analysis and MANOVA reveal that test familiarity and confidence significantly influence test experience. Higher-proficiency test-takers report greater self-efficacy and familiarity with the test format, suggesting that lower-proficiency students may benefit from additional support, such as pre-test instructions and practice materials. Cognitive strategy use, particularly planning and self-monitoring, is more pronounced among high-proficiency test-takers. Task 3 (integrated writing) relies more on planning and idea development, with graph synthesis posing challenges for lower-proficiency test-takers.

Teacher surveys indicate positive perceptions of both tasks in terms of design and effectiveness. The study highlights the BESTEP's alignment with Taiwan's EMI policy, integrating stakeholder insights into assessment development. Practical recommendations include targeted cognitive strategy training for lower-proficiency test-takers and further refinement of assessment practices to support EMI learners in academic writing. The study's findings have broader implications for EMI writing assessment and policy implementation.

Fan, J., Chen, I., & Knoch, U. (2024). Linking the BEST Test of English Proficiency (BESTEP) Writing Test to the Common European Framework of Reference (CEFR). Language Training and Testing Center.

This study aimed to link the BEST Test of English Proficiency (BESTEP) writing test to the Common European Framework of Reference (CEFR) using the CEFR Linking Manual's four stages: familiarisation, specification, standardisation, and validation. The BESTEP, developed by the Language Training and Testing Centre (LTTC) in Taiwan, assesses college students' readiness for academic English in Taiwan's tertiary education. The writing test includes three tasks: answering questions, expressing opinions, and writing an integrated essay.

Fifteen panellists participated, including six 'insiders' from Taipei and nine 'outsiders' from Australia. Insiders were familiar with the local context and the BESTEP, while outsiders had expertise in academic writing and the CEFR but little knowledge of the BESTEP. The study linked BESTEP score levels to CEFR levels and explored panellists' cognitive processes through a think-aloud study.

Three research questions were addressed: the relationship between BESTEP score levels and CEFR levels, the comparison of judgements between insiders and outsiders, and the panellists' mental processes during the linking. The Body of Work (BoW) method was used for standard setting, and multiple evidence types supported the validity of the linking results. The many-facets Rasch model (MFRM) analysed panellists' judgements, revealing that outsiders were generally more lenient than insiders, though all fit the Rasch model.

The think-aloud study identified themes in linking processes, strategies, and challenges. Both insiders and outsiders engaged in a dynamic, iterative linking process, focusing on similar performance aspects and strategies. The study's findings provide credible evidence for the alignment of BESTEP with CEFR levels and offer insights for future linking research.

Green, A., & Inoue, C. (2024). ***Relating the BESTEP Speaking Test to the Common European Framework of Reference for Languages*****. Language Training and Testing Center.**

This study investigates the alignment of the BESTEP Speaking Test with the Common European Framework of Reference for Languages (CEFR) to validate its effectiveness in assessing English proficiency for academic and professional contexts. The research follows the Council of Europe's recommended staged approach, including familiarization, specification, standardization, benchmarking, and validation. The BESTEP test, developed as part of Taiwan's Bilingual Education for Students in College (BEST) initiative, assesses English proficiency through structured speaking tasks and a rating system aligned with CEFR levels. A panel of language assessment experts conducted a standard-setting process using the Body of Work methodology, comparing BESTEP performance samples with established CEFR benchmarks. The findings confirm that the test's scoring system effectively differentiates proficiency levels. However, the findings suggested some discrepancies between the CEFR panelists' judgments and the original scores, which could be attributed to differences between the BESTEP scoring system and the Body of Work method used in this study. BESTEP scores are calculated by combining the scores awarded separately to each test part and include point deductions when candidates miss two or more questions. In contrast, the CEFR scales do not apply such penalties. The Body of Work method requires panelists to make an overall judgment of a test taker's level by balancing all available evidence. Further validation with a larger dataset is recommended to refine score interpretations and assess the test's impact on English-medium instruction (EMI) in Taiwan.

Lee, L. H. Y., Wu, R. Y. F., & Liao, C. H. Y. (2024). Evaluating scoring validity: Insights from the BESTEP English Proficiency Test in Taiwan [Paper presentation]. New Directions East Asia 2024, Bangkok, Thailand.

This study examines the scoring validity of the BESTEP, a language assessment developed for Taiwanese college students and launched in 2023. Scoring validity ensures that test scores are free from measurement errors and accurately reflect the abilities being measured, which is crucial for high-stakes English for Academic

Purposes (EAP) assessments. The BESTEP is aligned with higher education curricula and linked to the Common European Framework of Reference for Languages (CEFR), assessing students' academic English communicative skills, especially for English-medium instruction (EMI) courses.

Data from over 5,000 students' performances on the BESTEP speaking and writing tests were analyzed using the many-facet Rasch model (MFRM) to evaluate rater effects and test form difficulties. The results showed consistent rater severity and parallel test form difficulty. Additionally, analysis of covariance (ANCOVA) compared the performances of students from different academic disciplines (liberal arts, science/engineering, and biology/agriculture/medicine) and found no significant differences in their scores, indicating that BESTEP does not favor any specific majors.

These findings support the reliability and accuracy of BESTEP scores in measuring college students' academic English proficiency. The results provide higher education institutions with a tool to evaluate students' prerequisite skills before and after taking EMI courses. While qualitative approaches are also necessary to fully establish scoring validity, this quantitative investigation contributes fundamental evidence for scoring validity and enhances test transparency for stakeholders.

Lin, A. C., Wu, R. Y. F., & Wu, J. R. W. (2025). Aligning EAP teaching, learning, and assessment: A case study from Taiwan. International Journal of English for Academic Purposes: Research and Practice, 5(2), 241–281.

In 2021, Taiwan’s Ministry of Education launched the Programme on Bilingual Education for Students in College (BEST Programme) to enhance students’ English for Academic Purposes (EAP) skills and expand English Medium Instruction (EMI). Central to this initiative is the BEST Test of English Proficiency (BESTEP), designed to reflect language demands in university contexts. Performance descriptors developed across Common European Framework of Reference (CEFR) A2–B2 levels cover context, communication function, and linguistic performance. A mixed-methods sequential explanatory design was employed, entailing a survey of 805 students and 220 EAP/ EMI teachers, followed by interviews with a subset of participants. Analyses revealed broad agreement on the descriptors’ relevance, though EMI teachers differed on the

relevance of and student readiness for some descriptors. Students' perceived difficulty aligned with their proficiency, with tasks like synthesizing information identified as challenging. Findings highlight the value of stakeholder input in validating the BESTEP's descriptors and guiding EAP course design in support of academic progress.

Shin, S & Huang, S. (n.d.). ***Mapping BESTEP Listening and Reading Test Scores to CEFR Levels Using Test-Centered and Statistical Methods*****. [Ongoing project].**

This study aims to provide empirical evidence to align the BEST Test of English Proficiency (BESTEP) listening and reading section scores with the Common European Framework of Reference (CEFR; Council of Europe, 2001) levels by employing two standard-setting approaches: the Bookmark method (Karantonis & Sireci, 2006; Shin & Lidster, 2017) and statistical techniques, specifically hierarchical cluster analysis (HCA) (Shin & Lidster, 2017; Sireci et al., 1999) and latent class analysis (LCA) (Binici & Cuhadar, 2022; Brown, 2007). The cut scores generated by the Bookmark method will be compared across two groups of panelists, one from Indiana University and one from Taiwan, for cross validation. These scores will also be compared with cut scores derived from HCA and LCA, which provide data driven alternatives that do not rely on expert judgment. Together, these comparisons will offer a robust triangulation of methods.

Tono, Y. (n.d.). ***BESTEP learner corpora development project*****. Language Training and Testing Center [Ongoing project].**

This study aims to enhance the assessment of speaking proficiency by identifying criterial features for CEFR levels, focusing on grammar and vocabulary. The study has two main objectives: compiling a spoken learner corpus based on the BEST Test of English Proficiency (BESTEP) and investigating linguistic characteristics and error frequency to distinguish CEFR levels across BESTEP.

The study uses Weir's socio-cognitive framework, widely applied in global language proficiency tests like Cambridge English, IELTS, and TOEFL, to develop and validate the assessment criteria. The study will analyze language features such as grammar and vocabulary to validate assessment criteria, focusing on pronunciation,

fluency, coherence, lexical resource, grammatical range, and accuracy. The CEFR-informed English Profile for grammar and vocabulary will serve as a resource for this analysis.

The theoretical framework involves using learner corpora to examine scoring validity, with the Cambridge Learner Corpus (CLC) as a model. Error tagging in learner corpora will help identify error types and their frequency, contributing to language assessment development. Recent advancements in automated speech recognition (ASR) and automated speech scoring are also considered, despite existing challenges in AI scoring.

The research questions address the optimal corpus development process for using BESTEP results, identifying criterial features to distinguish CEFR levels, and reflecting various dimensions of language proficiency as defined by the CEFR. The study aims to improve CEFR level estimation and enhance the validity of speaking assessments.

Wu, J. R. W. (2023). ***Assessment in learning systems in Asia—Impeding, driving or improving?*** **[Panel discussion]. New Directions 2023, Hanoi, Vietnam.**

In this panel discussion, Dr. Jessica R. W. Wu shared how the LTTC, with the support of the Ministry of Education, developed and promoted BESTEP, a test designed specifically for university students, in the context of Taiwan's bilingual policy. Through needs analysis and the establishment of specific competency indicators and test objectives, as well as validity studies on stakeholder perceptions, curriculum relevance, and student performance, the LTTC has strengthened the interaction between teaching, learning, and assessment in higher education.

Wu, J. R. W., Huang, D. H. T., Hung, A. S. T., Lin, A. C. W., Chin, J. S., & Ke, A. S. H. (2024). ***A collaborative approach to examining BESTEP's impact on tertiary EAP in Taiwan*** **[Paper presentation]. 2024 Language Testing Research Colloquium (LTRC), Innsbruck, Austria.**

As part of the Bilingual Education for College Students (BEST) Program, the Ministry of Education in Taiwan has supported the development of the BEST Test of

English Proficiency (BESTEP). Aligned with the Common European Framework of Reference for Languages (CEFR) A2-B2 levels, BESTEP assesses English for Academic Purposes (EAP) abilities required at the tertiary level. This paper presents an ongoing validation study conducted jointly by BESTEP developers and two universities in Northern Taiwan. The study involves six EAP course designers and instructors and 200 students. Aiming to explore the connection between the BESTEP speaking and writing tests and EAP courses over two semesters, the study begins by analyzing and comparing course descriptions and test specifications, supplemented with classroom observations and teacher interviews. It then tracks changes in students' abilities using BESTEP as pre- and post-tests, alongside instructor evaluations from classroom assessments. The research highlights the vital integration of learning, teaching, and assessment, offering insights into means of improving language assessment practices and informing education system reforms. Additionally, it contributes to the conference theme of reforming language assessment systems by advocating for an effective approach to establishing a comprehensive learning system at the tertiary level in Taiwan.

Wu, J. R. W., Lin, A. C. W., & Chin, J. S. (2024). Enhancing learning, teaching, and assessment: The interconnection of BESTEP tests and EAP courses in Taiwan [Paper presentation]. 2024 International Conference on Bilingual Education, Taipei, Taiwan.

The BEST Test of English Proficiency (BESTEP) has been developed by the Language Training & Testing Center (LTTC) in support of Taiwan's Program on Bilingual Education for College Students (the BEST Program). Its goal is to bring about positive advancements in English for Academic Purposes (EAP) education at the tertiary level. This presentation centers on an ongoing validation study undertaken by LTTC in collaboration with two prominent universities in Taipei, involving a total of six classes and 200 students. Building upon our previous research, which primarily focused on creating performance descriptors that closely align with the requirements of university learning environments, the present study examines the relationship between BESTEP's speaking and writing tests and EAP courses. We conduct a comprehensive analysis that compares both the design of the courses and tests, while also tracking changes in students' speaking and writing abilities over two semesters.

This presentation is expected to shed light on the dynamic relationship between learning, teaching, and assessment in Taiwan's English language education. With a specific focus on the connection between BESTEP and EAP courses, this study validates the intended uses of BESTEP while emphasizing the importance of integrated approaches in EAP education. Findings from this study will significantly contribute to the conference theme of "Rethinking English Language Education at the Tertiary Level Under Taiwan's Bilingual Education policy," offering valuable insights and recommendations to educational stakeholders for a more effective approach.

Wu, J. R. W., Wu, R. Y. F. & Lin, A. C. W. (2023). *Co-constructing with stakeholders: The performance descriptors for an English productive skills test* [Paper presentation]. The 44th Language Testing Research Colloquium (LTRC), New York City, NY, United States.

In support of Taiwan's Bilingual Education policy, development of a standardized test which aims to assess and track college students' oral and written abilities expected in the English Medium Instruction (EMI) context is underway. A set of descriptors that

describe performance at three different levels, corresponding largely to CEFR A2-B2, has already been drafted for the test. In validating these descriptors and in justifying their intended usefulness, it is essential to explore the views held by teachers and students in relation to the descriptors in various dimensions, including how closely the described performance meets the demands of EMI learning contexts and how well students can perform in terms of the descriptors. This study employs a mixed-methods sequential explanatory design. Quantitative data comprise approximately 600 college students' and 30 English for Academic Purposes (EAP)/EMI teachers' responses to a survey based on the performance descriptors. Then, interviews with a small number of students and teachers are conducted. In addition to descriptive statistics, inferential statistical analyses are conducted to detect whether various background factors affect how respondents perceive students' EMI readiness. The study yields the following implications: 1. Co-constructing the validation of the performance descriptors with stakeholders will usefully inform the test development and ensure appropriate use of the test; 2. Understanding how EMI readiness is perceived among stakeholders will build common ground in defining EMI readiness; 3. Identifying students' difficulties in terms of the English language skills expected in an EMI context will be able to facilitate the improvement of EAP courses.

Wu, J. R. W., Wu, R. Y. F., Lin, A. C. W., & Chin, J. S. (2024). ***The BEST Test of English Proficiency (BESTEP): Contextual needs, realization, and future directions*** **[Panel discussion]. 2024 International Conference on Applied Linguistics & Language Teaching (ALLT), National Taiwan University of Science and Technology.**

This panel discussion provides an in-depth exploration of the BEST Test of English Proficiency (BESTEP), an English for Academic Purposes (EAP)-oriented assessment tool recently developed in support of Taiwan's bilingual education policy. Supported by the Ministry of Education and developed by the Language Training and Testing Center (LTTC), BESTEP adeptly addresses the needs of an increasingly globalized educational milieu in Taiwan's higher education, while maintaining alignment with the CEFR framework. Furthermore, BESTEP recognizes its pivotal role in fostering the language proficiency of EFL learners in Taiwan and employs innovative measures aimed at facilitating the integrations of assessment, instruction, and learning.

The 1st paper presented by Jessica R. W. Wu will delve into the origins of BESTEP, with a focus on the specific contextual requirements and policy influences that shaped its development. The 2nd paper presented by Rachel Y. F. Wu will examine the intricacies of the design and implementation of BESTEP, highlighting its alignment with EAP skills crucial for university-level education and its integration with principles of learning-oriented assessment. The 3rd paper presented by Joyce S. Chin will explore the BESTEP's potential impact on EAP instruction and learning at the tertiary level, drawing the interactive connections to broader socio-cultural dynamics. The 4th paper presented by Anita C. W. Lin will showcase the resources and support mechanisms designed for different BESTEP's stakeholders and investigate their implications for fostering a positive feedback loop among learning, teaching, and assessment. The panel will conclude with a discussion on potential avenues for advancing BESTEP within a bilingual educational environment.

Wu, J. R. W. & Wu, R. Y. F., & Liu Anne. (2022). Developing an EAP test for college students in support of Taiwan's bilingual 2030 policy [Paper presentation]. Academic Forum on English Language Testing in Asia (AFELTA), Bangkok, Thailand.

In 2021, Taiwan's Ministry of Education (MOE) introduced "The Program on Bilingual Education for Students in College" (the BEST project) to promote bilingual education and to improve the English proficiency of university students. To provide a tool to track the KPIs of the project and create a more motivating English learning environment, the MOE commissioned the Language Training and Testing Center (LTTC) to develop a standardized test which aligns with the learning objectives of English education at the tertiary level in Taiwan and, moreover, links with the National Curriculum Guidelines for 12-year Basic Education. Different from current English proficiency tests, this new test is designed to assess college students' English proficiency for academic purposes (EAP) in the context of English-medium instruction (EMI) in Taiwan. In this presentation, we focus on the development of the speaking and writing sub-tests, including major test features, rating scales, and pretest results.

Appendix 2. Testimonial from Professor Barry O'Sullivan OBE

Testimonial

The BESTEP is a quality assessment tool, thoughtfully aligned with both Taiwan's higher education context and international frameworks such as the CEFR. Remarkably, despite being launched only three years ago, the BESTEP has already made a significant positive impact on English teaching and learning at the tertiary level. In many cases, such key achievements are typically observed only after much longer periods of use. The success of BESTEP, made possible through LTTC's dedicated efforts, serves as a compelling exemplar of test glocalization for other countries with similar needs. With the solid groundwork that BESTEP has established, I am confident that it will continue to grow and thrive in the years to come.

Professor Barry O'Sullivan OBE
December 20th, 2024