p-ISSN 2252-7427, e-ISSN 2541-3635
https://journal.uniku.ac.id/index.php/IEFLJ/index

p-ISSN 2252-7427 e-ISSN 2541-3635
Indonesian EFL Journal
AISEE
The Association of Indonesia Scholars of English Education

A QUANTITATIVE CROSSWALK BETWEEN TOEFL ITP SCORES AND CEFR LEVELS: AN OUTCOME-BASED APPROACH FOR ELT IN HIGHER EDUCATION

Erna Nurkholida[1*], Ulin Ni'mah Rodliya Ulfah[2], Rafidah binti Abdullah[3], Titik Setiasih[4]
[1]*English Department, UIN Syekh Wasil Kediri, Indonesia.*
[2]*Faculty of Islamic Economics and Business, UIN Syekh Wasil Kediri, Indonesia*
[3]*Faculty of Arabic Language, University of Islam Sultan Sharif Ali, Brunei Darussalam*
[4]*Faculty of Tarbiyah and Teacher Training, UIN Syekh Wasil Kediri, Indonesia*

Corresponding author's email: ernanurkholida@uinkediri.ac.id

APA Citation: Nurkholida, E., Ulfah, U. N. R., Abdullah, R., & Setiasih, T. (2026). A quantitative crosswalk between TOEFL ITP scores and CEFR levels: An outcome-based approach for ELT in higher education. *Indonesian EFL Joufrnal, 12*(2), 543-554. https://doi.org/10.25134/ieflj.v12i2.138

Received: 12-01-2026 Accepted: 25-05-2026 Published: 29-05-2026

Abstract: As the Common European Framework of Reference for Languages (CEFR) gains prominence as a global benchmark, higher education institutions require precise and transparent methods to interpret institutional English proficiency test scores. However, the absence of empirical alignment between TOEFL ITP scores and CEFR levels poses challenges for outcome-based academic environments, particularly in relation to placement, certification, and policy decisions. This study aims to develop a quantitative crosswalk between TOEFL ITP scores and CEFR proficiency levels within an Outcome-Based Education (OBE) framework. Using a quantitative cross-sectional design, data were collected from 90 students in the seventh-semester university who completed a preparatory course and subsequently took an online TOEFL ITP prediction test via the institutional Learning Management System. The relationship between TOEFL ITP scores and CEFR classifications was analyzed using descriptive statistics, Pearson correlation, and linear regression. The findings reveal a strong and statistically significant positive correlation between TOEFL ITP scores and CEFR levels ($r = .78$, $p < .001$). Regression analysis indicates that CEFR levels explain 60.8% of the variance in TOEFL ITP scores ($R^2 = .608$), with an average increase of approximately 90 points per CEFR level. Most participants were classified at the B1 level, indicating an intermediate proficiency typical of students preparing for academic contexts. These findings provide empirical support for aligning TOEFL ITP scores with CEFR descriptors and highlight the potential of an outcome-based approach to enhance transparency, accountability, and curriculum-aligned assessment practices in English Language Teaching (ELT) in higher education.

Keywords: *TOEFL ITP; CEFR; language proficiency assessment.*

INTRODUCTION

Higher education institutions increasingly require dependable and transparent measures of English proficiency to inform admissions, placement, and progression decisions such as admission, placement, and program progression. Among institutional tests, the TOEFL ITP is widely used due to its accessibility and practicality. At the same time, the Common European Framework of Reference for Languages (CEFR) has been internationally recognized as a standardized framework for describing language proficiency across levels A1 to C2. However, the lack of empirically validated alignment between TOEFL ITP scores and CEFR levels creates challenges not only for institutional decision-making but also for English Language Teaching (ELT) practices, particularly in placing students appropriately, designing level-based instruction, and monitoring learning progression. The absence of an empirically verified crosswalk continues to create ambiguity for stakeholders, including teachers and curriculum designers (Educational Testing Service, 2023; Papageorgiou et al., 2021).

A substantial body of research on score linking has focused on large-scale assessments such as

TOEFL iBT and IELTS, where statistically validated alignment methods have been widely applied; however, comparable empirical work on TOEFL ITP remains limited. Existing mappings are often based on expert judgment rather than rigorous statistical validation (Educational Testing Service, 2023). Moreover, methodological challenges arise due to differences in construct coverage: TOEFL ITP emphasizes receptive academic skills such as listening and reading, whereas CEFR reflects communicative competence across multiple language skills (Council of Europe, 2020; North, 2014). As a result, the absence of a statistically grounded crosswalk limits the pedagogical usefulness of TOEFL ITP scores in ELT contexts, where teachers require interpretable proficiency benchmarks to inform instruction and assessment. Scholars emphasize that statistically supported linking studies are essential to ensure fairness, comparability, and interpretive validity in language assessment (Kane, 2013; Kolen & Brennan, 2020).

While the Educational Testing Service (2023) provides indicative mappings between TOEFL ITP scores and CEFR levels, these are primarily based on expert judgment and internal standard-setting procedures, with limited publicly available empirical validation. As a result, the underlying statistical relationships between TOEFL ITP scores and CEFR proficiency descriptors remain insufficiently examined in independent research contexts.

This study addresses this methodological gap by applying correlation and regression analyses to empirically model the relationship between TOEFL ITP scores and CEFR levels using observed learner data, thereby providing a transparent and data-driven alternative to existing concordance tables.

Previous research on score linking has produced several established approaches to aligning language tests with the CEFR, particularly through standardized procedures outlined in the *Manual for Relating Language Examinations to the CEFR* (Council of Europe, 2009, 2020). These approaches emphasize empirical validation using statistical methods such as Equi percentile linking, regression analysis, and standard-setting techniques (Karantonis & Sireci, 2022; Kolen & Brennan, 2020). the Shapiro–Wilk test ($W = 0.97$, $p = .084$)

Previous score-linking studies have reported moderate to strong correlations between standardized English proficiency tests and CEFR levels. For example, Papageorgiou et al. (2021) identified correlations ranging from 0.72 to 0.85 in linking TOEFL iBT scores with CEFR proficiency levels. Similarly, Ginther and Yan (2018) reported correlations of approximately 0.75 between TOEFL ITP scores and external measures of language proficiency. Earlier validation studies have also shown coefficients of determination ranging from 0.55 to 0.60, indicating that standardized test scores can explain a meaningful proportion of variance in CEFR-related proficiency classification. These findings collectively support the feasibility of aligning TOEFL-based assessments with CEFR frameworks.

However, the existing body of research also reveals several limitations. First, most studies focus on high-stakes, four-skill assessments, leaving institutional tests such as TOEFL ITP underexplored. Second, while reliability and validity evidence are well-documented for TOEFL iBT and IELTS, similar empirical investigations into TOEFL ITP score distributions and their alignment with CEFR levels remain limited. Third, many available TOEFL ITP–CEFR mappings rely on expert judgment or approximate concordance tables rather than rigorous statistical modeling. These gaps highlight the need for a systematic, data-driven approach to linking TOEFL ITP scores with CEFR proficiency levels.

A key methodological gap remains in the lack of statistically validated alignment between TOEFL ITP scores and CEFR proficiency levels, as existing mappings—particularly those provided by the Educational Testing Service—are largely based on expert judgment rather than transparent empirical modeling.

Therefore, a statistically grounded approach—such as correlation analysis and regression modeling—provides a more robust basis for linking scores and addressing existing limitations in prior studies. This study addresses this gap by developing a quantitative crosswalk between TOEFL ITP scores and CEFR proficiency levels using correlation analysis and regression modelling with the aim of supporting accurate interpretation and curriculum-aligned ELT practices in higher education. Beyond its measurement contribution, this study explicitly situates score alignment within ELT practices. The resulting mapping is expected to support (1) classroom placement decisions by aligning test scores with CEFR-based proficiency levels, (2) syllabus design through clearer identification of target competencies at each level, and (3) learning progression by enabling instructors to monitor

p-ISSN 2252-7427, e-ISSN 2541-3635
https://journal.uniku.ac.id/index.php/IEFLJ/index

students' development using standardized descriptors. In this way, the study contributes not only to language assessment research but also to pedagogical decision-making in ELT contexts.

Within this context, Outcome-Based Education (OBE) provides a relevant pedagogical framework. OBE emphasizes clearly defined and measurable learning outcomes and requires alignment among curriculum, instruction, and assessment (Spady, 1994; Biggs & Tang, 2022). Research indicates that effective implementation of OBE depends on the coherence between teaching practices and assessment systems (Schuwirth & van der Vleuten, 2023), as well as on educators' ability to design outcome-oriented learning experiences (Harden, 2019). In language education, this implies that assessments should reflect learners' communicative abilities in observable and measurable ways (Leung & Scarino, 2021).

Aligning TOEFL ITP scores with CEFR levels operationalizes OBE by transforming test scores into meaningful indicators of learning outcomes. CEFR descriptors, which are formulated as observable performance benchmarks, align naturally with OBE principles (Council of Europe, 2020).

However, many existing crosswalk studies emphasize psychometric validation without explicitly linking results to curriculum implementation or instructional practices (Islam & Sood, 2023). Without such integration, score alignment risks remaining a technical exercise rather than a pedagogically meaningful tool. To address this limitation, the present study embeds TOEFL ITP–CEFR alignment within an OBE-oriented ELT framework. By interpreting TOEFL ITP scores in relation to CEFR-based learning outcomes, the study enables their application in placement decisions, syllabus development, and learning progression tracking. This approach enhances the transparency, accountability, and educational relevance of language assessment in higher education (Mistamiruddin & Mohamad Nasri, 2024).

In particular, limited attention has been given to examining TOEFL ITP score distributions, internal consistency, and validity arguments in relation to CEFR descriptors, which are essential components in establishing defensible score interpretations.

Therefore, this study contributes to the existing literature by providing an empirically derived TOEFL ITP–CEFR score alignment based on observed learner data, offering a transparent and data-driven complement to existing expert-judgment mappings such as those provided by ETS. The novelty of this study lies in the application of established statistical linking techniques to TOEFL ITP within a real instructional context, combined with the explicit integration of the resulting alignment into an Outcome-Based Education (OBE) framework. While the statistical techniques employed are well-established, their application to TOEFL ITP in a localized higher education context addresses a gap in the current literature. Furthermore, the study extends beyond psychometric analysis by interpreting the resulting alignment within an Outcome-Based Education (OBE) framework, thereby illustrating its potential use in placement decisions, syllabus design, and monitoring of learning progression. It should be noted, however, that the findings are context-specific and intended to provide exploratory rather than broadly generalizable insights.The research question guiding this study is: How can TOEFL ITP scores be statistically aligned with CEFR proficiency levels to support accurate interpretation and pedagogically meaningful application in higher education ELT contexts?

METHOD

This study employed a quantitative cross-sectional design to develop a statistical crosswalk between TOEFL ITP scores and CEFR proficiency levels. CEFR proficiency levels were assigned using the Bookmark standard-setting method with three expert judges, following the guidelines of the Council of Europe (2020). Score-linking research commonly uses quantitative methods to empirically connect standardized test scores with external proficiency benchmarks (Kolen & Brennan, 2020; Papageorgiou et al., 2021). Regression modeling and correlational analyses were selected as the primary analytical techniques because they provide robust and interpretable mappings across assessment scales (Papageorgiou et al., 2021; Powers & Schedl, 2021). The TOEFL ITP prediction test demonstrated acceptable internal consistency reliability in the present sample (KR-20 = .84). Contemporary validation frameworks further emphasize the importance of empirical evidence in supporting score interpretation and use (Chapelle et al., 2021; Kane, 2013).

The population of this study consisted of 150 seventh-semester students at UIN Syekh Wasil Kediri from various academic disciplines. Based on sample size recommendations for quantitative research (Creswell & Creswell, 2018), a minimum of 90 participants was determined to be sufficient

to achieve a 5% margin of error. A total of 90 students were selected using random sampling techniques. Only students who completed the TOEFL ITP prediction test and had complete score data were included in the final analysis. The selection of seventh-semester students was intentional, as their proficiency levels are critical for graduation requirements, placement decisions, and institutional evaluation in higher education contexts.

The researcher administered an online TOEFL ITP prediction test through the institutional Learning Management System (LMS) to collect the data. The test assessed listening comprehension, structure and written expression, and reading comprehension.

CEFR proficiency levels were assigned based on score interpretation procedures aligned with CEFR descriptors and institutional benchmarks. TOEFL ITP score ranges were mapped onto CEFR levels using a regression-based crosswalk approach supported by cut-score determination. In this study, CEFR levels function as observable proficiency indicators reflecting learners' communicative abilities as described in CEFR descriptors (Council of Europe, 2020; North, 2014). Thus, CEFR classification serves as a proxy for outcome-based proficiency levels.

The data were analyzed using descriptive statistics, Pearson product-moment correlation, regression modelling, and cut-score analysis. Descriptive statistics were used to summarize score distributions and identify general performance trends. For statistical analysis, CEFR levels were coded ordinally as A1 = 1, A2 = 2, B1 = 3, B2 = 4, and C1 = 5. This coding assumes ordinal progression but does not imply equal intervals between proficiency levels, consistent with standard practice in CEFR alignment studies (Papageorgiou et al., 2021).

Prior to conducting the correlation and regression analyses, several statistical assumptions were assessed. The normality of TOEFL ITP scores was examined using the Shapiro–Wilk test ($W = 0.97$, $p = .084$), indicating no significant deviation from normality. Linearity was confirmed through visual inspection of the scatterplot. In addition, residual analysis revealed no violations of homoscedasticity or independence assumptions.

Pearson correlation analysis was conducted to examine the strength and direction of the relationship between TOEFL ITP scores and CEFR levels (Field, 2022).

Regression analysis was then applied to predict CEFR levels based on TOEFL ITP scores, following established score-linking methodologies (Kolen & Brennan, 2014; Papageorgiou et al., 2021). Cut-score determination was used to define score intervals corresponding to CEFR levels, ensuring interpretability and alignment with proficiency benchmarks.

Within an Outcome-Based Education (OBE) framework, CEFR levels were treated as measurable learning outcomes representing students' communicative proficiency at different stages. This approach is consistent with OBE principles, which emphasize clearly defined and assessable learning outcomes as the basis for curriculum and assessment design (Spady, 1994; Biggs & Tang, 2011).

The TOEFL ITP–CEFR crosswalk developed in this study enables test scores to be interpreted in terms of outcome attainment, thereby supporting: placement decisions (assigning students to appropriate proficiency levels), curriculum alignment (matching instructional content with CEFR-based competencies), and learning progression monitoring (tracking development across proficiency levels).

By embedding statistical alignment within an outcome-based framework, this study ensures that score interpretation extends beyond psychometric analysis to support pedagogically meaningful applications in English Language Teaching (ELT).

RESULTS AND DISCUSSION

This section presents and discusses the TOEFL ITP scores obtained by students who completed a TOEFL preparation course. The goal of the course was to help students learn how to handle the test's listening, structuring, and reading components as well as to familiarize them with its format.

The researcher administered the test online via the school's Learning Management System (LMS), and each class had about 25 pupils. Table 1 provides a broad overview of the students' English proficiency levels by summarizing the mean scores attained following the preparation course.

A quantitative correlational approach was used in the study, which is suitable for investigating connections between quantifiable variables (Ary et al., 2023; Creswell & Creswell, 2023). Pearson product-moment correlation analysis was used to investigate the relationship between performance on individual test portions and the total TOEFL ITP score. The direction and strength of linear correlations between continuous variables are frequently ascertained using this statistical method (Field, 2022).

p-ISSN 2252-7427, e-ISSN 2541-3635
https://journal.uniku.ac.id/index.php/IEFLJ/index

Table 1. *Descriptive statistic*

Statistic	Value
Count (N)	90
Mean (Average)	398.59
Median	381.50
Standard Deviation (σ)	68.61

The relatively high standard deviation (SD = 68.61) indicates substantial variability in students' proficiency levels. Such variability is expected in standardized assessments and supports meaningful score interpretation, as emphasized in the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2020). This suggests that differences in learners' prior proficiency and exposure to English continue to influence performance despite a shared instructional context.

This variability is particularly important for the present study, as it enables meaningful differentiation across CEFR proficiency levels and supports subsequent correlation and regression analyses.

Cut score table (TOEFL ITP → CEFR)

Table 2. *Cut score table*

CEFR Level	Min Score	Max Score	Frequency (N)
A1	177	177	1
A2	313	330	8
B1	337	470	70
B2	463	533	9
C1	550	647	2

The distribution of students' TOEFL ITP scores transferred onto CEFR skill levels is shown in Table 2. According to the results, most students (N = 70) had intermediate English ability, with a B1 classification. The dominance of the B1 proficiency level has important implications for teaching strategies. As B1 represents an intermediate threshold where learners can handle familiar topics but still struggle with complex academic discourse, instructional practices should focus on scaffolding learners toward higher-order communicative competence, particularly in productive skills such as speaking and writing. This includes task-based learning, integrated skills instruction, and increased exposure to authentic academic texts. The concentration at the B1 level also suggests that short-term TOEFL preparation courses may be more effective in consolidating intermediate proficiency rather than promoting rapid advancement to higher CEFR bands, highlighting the need for sustained and curriculum-integrated language development.

Within an Outcome-Based Education (OBE) framework, the TOEFL ITP–CEFR crosswalk operationalizes learning outcomes by providing empirically grounded proficiency benchmarks. The predominance of students at the B1 level suggests that the TOEFL preparation course was primarily aligned with intermediate proficiency outcomes. However, the relatively limited number of students achieving B2 and C1 levels indicates that curricular adjustments and additional instructional scaffolding may be necessary to support progression toward higher proficiency bands.

Only a very small percentage of students attained the advanced C1 level (N = 2), while smaller percentages were classified at the A2 (N = 8) and B2 (N = 9) levels. Very basic proficiency was indicated by one student's placement at the A1 level. According to this distribution, most participants showed functional academic English skills rather than advanced proficiency, even though they had taken a TOEFL preparatory course. From the standpoint of learning outcomes, this pattern is consistent with the usual development seen in short-term language preparation programs, which typically result in the consolidation of intermediate competence rather than the quick ascent to higher CEFR levels (Biggs et al., 2022).

Pearson correlation analysis was used to further investigate the connection between CEFR categories and TOEFL ITP results. Higher TOEFL ITP performance is regularly linked to higher CEFR proficiency categories, according to the results, which showed a strong positive correlation between overall TOEFL ITP scores and assigned CEFR levels. This methodical progression from A1 to C1 validates that TOEFL ITP scores serve as significant markers of general language competency within the framework of the CEFR and supports the validity of the quantitative score mapping.

The work of Zheng and Yan highlights that statistically supported score interpretations improve the comparability and transparency of extensive language tests across competency frameworks, is in line with these findings. Nonetheless, the clustering of results within the B1 proficiency band reveals a significant measurement-related pattern. Many students maintained their CEFR level despite numerical gains in their TOEFL ITP scores, suggesting that score improvement does not always correspond to advancement in a category level.

Since CEFR descriptors define competency in terms of functional language use rather than

specific score distinctions, this phenomenon reflects their intrinsically broad nature. Schu Wirth (2023) contends, from the standpoint of assessment theory, that categorical proficiency frameworks invariably reduce score variability within bands, possibly hiding significant individual variations.

Biggs et al. (2022) also point out that learning outcomes frameworks give threshold attainment more weight than incremental score increases, which explains why students might show quantifiable progress without going over a CEFR boundary. When combined, these results highlight the necessity for careful interpretation while also confirming the TOEFL ITP–CEFR mapping's general validity.

The categorical character of CEFR classifications means that TOEFL ITP results must be used carefully in high-stakes situations like placement, certification, or advancement decisions, even though Pearson correlation analysis shows a substantial alignment between TOEFL ITP scores and CEFR levels. Therefore, to promote equitable and pedagogically sound decision-making, institutions are advised to supplement CEFR-referenced score interpretations with additional performance information.

Table 3. *Correlation table*

Metric	Value	Interpretation
Pearson Correlation	0.780	Strong Positive Correlation
p-value	1.36 \times 10^ {-19}	Statistically Significant ($p < 0.05$)
N (Sample Size)	90	

Note. The Pearson correlation coefficient yielded a 95% confidence interval ranging from 0.68 to 0.85, indicating a stable and reliable positive relationship between CEFR level and TOEFL score.

The TOEFL ITP scores and CEFR proficiency levels were found to be strongly and statistically significantly correlated by the Pearson product-moment correlation analysis. According to Table 3, there is a strong positive link between higher TOEFL ITP scores and higher CEFR proficiency classes, as indicated by the correlation coefficient ($r = .78$, $p < .001$). This outcome demonstrates that TOEFL ITP performance reflects advancement throughout the CEFR proficiency continuum. Further evidence that the observed association is unlikely to have happened by chance is provided by the incredibly small p-value (1.36×10^{-19}), which is much below the traditional significance threshold of.05. The investigation offers strong empirical support and significant statistical power for the validity of the suggested quantitative crosswalk, with a sample size of 90 participants.

In score-linking and alignment studies between external proficiency frameworks and standardized tests, strong statistical connections are often considered to be crucial evidence (Papageorgiou et al., 2024; Council of Europe, 2020). The extent of the link shows that TOEFL ITP scores act as a valid quantitative indicator of CEFR competence levels, particularly within the intermediate range where most participants were classified.

A substantial correlation does not, however, indicate that improvements in numerical scores and advances in categorical proficiency are equivalent. According to recent validation study, proficiency frameworks like the CEFR do not represent equal-interval measuring scales but rather large performance bands (Kane, 2021; Chappelle, Enright, & Jamieson, 2021).

As a result, TOEFL ITP score variance might happen within the same CEFR level without always leading to category advancement. These results emphasize how crucial it is to interpret TOEFL ITP results carefully when they are used to make important academic or institutional decisions like placement, certification, or promotion.

In order to ensure justice and responsible score utilization, current talks on the implementation of the CEFR emphasize that contextual and pedagogical considerations should be added to statistical alignment (Little & Figueras, 2022; Council of Europe, 2020). Despite recognizing the inherent drawbacks of categorical proficiency frameworks, the Pearson correlation study shows a significant alignment between TOEFL ITP results and CEFR proficiency levels overall.

Scatterplot of CEFR levels and TOEFL ITP scores with regression line

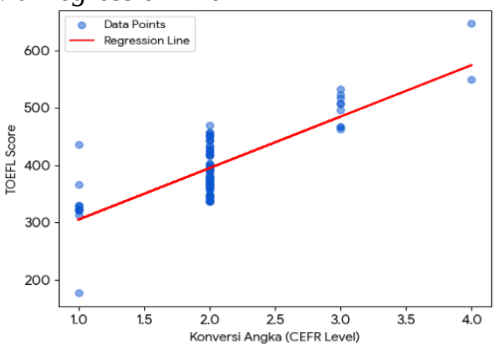


Figure 1. *Relationship between CEFR level and TOEFL score*

Figure 1. Scatterplot showing the relationship between CEFR proficiency level and TOEFL

p-ISSN 2252-7427, e-ISSN 2541-3635
https://journal.uniku.ac.id/index.php/IEFLJ/index

scores. The red line represents the fitted linear regression model. It illustrates a positive linear relationship between CEFR level and TOEFL score, indicating that higher CEFR proficiency levels are associated with higher TOEFL performance.

Convincing proof that TOEFL ITP scores are significantly correlated with CEFR proficiency levels can be seen in the regression and correlation results. According to the regression model's steady rising trend, students' placement along the CEFR continuum appears to improve as their TOEFL ITP scores rise. This pattern shows a consistent and comprehensible alignment between a standardized exam and an external proficiency framework rather than suggesting precise score equivalency. Systematic correspondence is more significant than perfect numerical matching, according to similar findings from extensive validation and linkage research, including studies employing MELAB and other competency benchmarks (Hidri, 2020).

However, the findings indicate that, especially at intermediate levels, students may still have disparate TOEFL ITP scores even if they are in the same CEFR band. The nature of the CEFR itself is reflected in this variety, which is not problematic. According to studies on the CEFR, proficiency levels are not set measurement intervals but rather are intended as broad communicative categories (Sirakoern Tangsakul, 2024).

It makes sense for students with marginally differing test scores to stay in the same band since CEFR descriptors place a high value on the use of functional language. This explains why the regression analysis shows a substantial overall relationship in addition to within-level variability.

These results are strongly related to the constructive alignment principle from a teaching standpoint, which emphasizes consistency between instructional activities, assessment procedures, and intended learning outcomes (Biggs & Tang, 2022).

The robust statistical correlation suggests that the TOEFL preparation course was widely in line with CEFR-based competency development rather than only test-oriented. To secure effective acceptance, recent talks on the application of the CEFR in various national contexts also stress how crucial it is to harmonize curriculum frameworks, assessment methods, and institutional policies (Maryo, 2021; Hazar, 2021).

These studies highlight that without alignment, CEFR risks becoming a symbolic reference rather than a functional framework guiding learning outcomes and assessment practices. Taken together, the findings support the use of TOEFL ITP scores as indicators of CEFR-referenced language development. Nevertheless, careful interpretation remains essential, particularly when such scores are used for high-stakes academic or institutional decision-making.

Regression line of CEFR conversion with TOEFL score

The evidence for a significant correlation between CEFR proficiency levels and TOEFL ITP scores is further supported by the linear regression analysis. According to the regression slope (m = 89.81), TOEFL ITP scores typically improve by over 90 points for every CEFR level increase. This difference is not insignificant. Significant increase in language proficiency is shown, and the idea that test results should show distinct developmental progression across levels is supported (Rao, 2020; Spady & Wu, 2024).

Table 4. *Regression line of CEFR conversion with TOEFL score*

Metric	Value	Interpretation
Slope (m)	89.81	For every one-level increase in CEFR proficiency, the TOEFL score increases by approximately 90 points.
Intercept (c)	215.48	Represents the theoretical baseline TOEFL score when the CEFR conversion value is zero.
R-squared (R^2)	0.608	Indicates that 60.8% of the variability in TOEFL scores is explained by CEFR level.
Standard Error	7.69	Represents the average deviation of the observed values from the regression line.
P-value	< 0.001	Indicates that the relationship between CEFR level and TOEFL score is statistically significant.

Note. The regression slope demonstrated a 95% confidence interval ranging from 76.20 to 103.42, while the intercept showed a confidence interval ranging from 190.11 to 240.85. These intervals indicate that the regression estimates are statistically reliable.

The results of the correlation analysis are presented in Table 3, while the regression model is summarized in Table 4.

A significant amount of the score variation can be explained by the model. With a R2 value of.608,

CEFR proficiency levels account for roughly 60.8% of the variation in TOEFL ITP scores. This indicates a significant structural relationship in score-linking research, particularly when a continuous numerical scale like the TOEFL ITP relates to a category framework like the CEFR (Babu & Ray, 2023; Kolen & Brennan, 2022).

Since each system captures slightly different characteristics of linguistic ability, complete prediction is neither theoretically nor practically predicted in cross-framework alignment (Sinha, 2021). The majority of observed TOEFL ITP scores also fall near the regression line, as indicated by the comparatively small standard error (7.69).

This implies that the suggested crosswalk is both stable and somewhat accurate. CEFR mappings can attain reliable levels of accuracy when measurement error is controlled, according to similar regression-based linking studies (Akramy, 2021; Papageorgiou et al., 2024). The observed association is statistically significant and unlikely to be the result of chance, as further confirmed by the very significant p-value (<.001).

These results are consistent with Kane's (2021) argument-based approach to validation, which stresses that score interpretations need to be backed up by both theoretical justification and statistical data. Nonetheless, there is always a degree of ambiguity when relating general proficiency frameworks to standardized tests (Sheikh Ul Islam, 2023). Students at the same level may nonetheless exhibit notable variances in their TOEFL ITP scores since CEFR levels indicate broad communicative bands (Zheng & Yan, 2025).

In addition to reminding us to carefully interpret score mappings, particularly in high-stakes situations like placement, certification, or institutional policy decisions, the regression results collectively offer strong empirical support for the TOEFL ITP–CEFR crosswalk.

ITP–CEFR cut-scores with ETS (2023) mappings

Table 5. *Comparison of TOEFL*

CEFR Level	Current Study (Min Score)	ETS (2023) Indicative Range	Difference	Interpretation
A2	313	310–340	+3	Comparable
B1	337	340–460	−3	Slightly lower threshold
B2	463	470–540	−7	Slightly lower threshold
C1	550	550–620	0	Identical

Note. Differences may reflect sample characteristics, institutional context, and methodological variation in score-linking procedures.

The correlation coefficient ($r = 0.78$) observed in the current study is comparable to those reported in similar score-linking studies. Papageorgiou et al. (2021) reported correlations ranging from 0.72 to 0.85 when linking TOEFL iBT scores to CEFR levels, while Ginther and Yan (2018) identified correlations of approximately 0.75 between TOEFL ITP scores and external proficiency measures.

In addition, the coefficient of determination obtained in the present study ($R^2 = 0.608$) indicates that the regression model explains a substantial proportion of score variance, slightly exceeding the 0.55–0.60 range reported in several earlier validation studies. This difference may be attributed to the relatively homogeneous sample characteristics and institutional context of the current study.

Beyond statistical alignment, these findings carry important pedagogical implications for English Language Teaching (ELT). The strong relationship between TOEFL ITP scores and CEFR levels suggests that test results can be meaningfully interpreted as indicators of learners' proficiency development. In classroom contexts, this enables instructors to use TOEFL ITP scores not merely as numerical outcomes but as diagnostic tools for placement and instructional planning. Students classified at similar CEFR levels can be grouped more effectively, allowing teachers to design level-appropriate learning activities and materials aligned with CEFR descriptors (Council of Europe, 2020; Biggs & Tang, 2011).

The findings also highlight the need for stronger curriculum alignment in ELT programs. The ability to map TOEFL ITP scores onto CEFR levels provides a practical framework for designing CEFR-aligned syllabi, where learning objectives, instructional materials, and assessment practices are coherently structured around proficiency descriptors. Without such alignment, standardized test results risk being used solely for administrative purposes rather than informing teaching and learning processes. As emphasized in constructive alignment theory, effective instruction requires consistency between intended learning outcomes, teaching activities, and assessment methods (Biggs & Tang, 2011).

In practical terms, the TOEFL ITP–CEFR crosswalk developed in this study should be viewed as a supporting tool for pedagogical

p-ISSN 2252-7427, e-ISSN 2541-3635
https://journal.uniku.ac.id/index.php/IEFLJ/index

decision-making, rather than a definitive measure of proficiency. When used alongside classroom-based assessment and continuous evaluation, it can enhance transparency, improve placement accuracy, and support outcome-based curriculum design in higher education ELT contexts.

From an Outcome-Based Education perspective, this study contributes by operationalizing CEFR levels as measurable learning outcomes that can inform assessment interpretation. Rather than viewing test scores as isolated numerical indicators, the TOEFL ITP–CEFR crosswalk enables institutions to interpret results in terms of progressive competence levels. However, the findings also suggest that progression across CEFR levels requires more than incremental score gains, reinforcing the importance of aligning assessment with clearly defined learning outcomes and communicative performance benchmarks (Spady, 1994; Biggs & Tang, 2011).

In terms of practical implications, this study offers several contributions to ELT practice in higher education. For teachers, the crosswalk can function as a diagnostic tool to support more accurate classroom placement and to guide the selection of level-appropriate instructional strategies. For curriculum designers**,** the alignment provides a basis for developing CEFR-informed syllabi that link learning objectives, teaching activities, and assessment practices. For language centers and institutions**,** the crosswalk enhances transparency and accountability in placement, certification, and program evaluation by providing a more interpretable framework for understanding TOEFL ITP scores.

Despite the strong statistical relationship observed, it is important to acknowledge the limitations of TOEFL ITP as a proficiency measure. The test primarily assesses receptive skills (listening and reading) and language structure, with limited representation of productive and interactive communicative abilities. In contrast, CEFR descriptors emphasize functional language use across all four skills, including speaking and writing (Council of Europe, 2020). As a result, the TOEFL ITP–CEFR crosswalk may not fully capture learners' communicative competence, particularly at higher proficiency levels. Therefore, reliance on TOEFL ITP scores alone may lead to incomplete interpretations of learners' abilities and should ideally be complemented with performance-based assessments.

In addition, the regression and score-linking model developed in this study was derived from a single institutional sample and was not cross-validated using an independent dataset. As a result, the generalizability of the proposed TOEFL ITP–CEFR crosswalk should be interpreted cautiously. Future research is recommended to conduct cross-validation with larger and more diverse samples to strengthen the stability and applicability of the model across different educational contexts.

CONCLUSION

This study examined whether TOEFL ITP scores can be meaningfully interpreted in relation to CEFR proficiency levels within an Outcome-Based Education (OBE) framework. The findings demonstrate a strong and statistically significant relationship between TOEFL ITP scores and CEFR levels, supported by both correlation ($r = .78$) and regression analysis ($R^2 = .608$ TOEFL ITP scores can serve as indicators of learners' receptive proficiency when interpreted through CEFR descriptors. The predominance of learners at the B1 level reflects a typical intermediate proficiency profile in academic preparatory contexts, suggesting that short-term preparation programs tend to consolidate rather than substantially elevate proficiency levels.

However, the applicability of the proposed crosswalk should be interpreted with caution. The TOEFL ITP primarily assesses receptive skills and structural knowledge, whereas CEFR emphasizes broader communicative competence across productive and interactive skills (Council of Europe, 2020). As a result, the alignment may not fully capture learners' overall communicative ability. In addition, the study is based on a single institutional sample, which may limit the generalizability of the findings to other educational contexts. These limitations highlight the need for complementary assessment methods and further validation studies.

Overall, this study contributes to the growing body of research on score linking and proficiency alignment by demonstrating that statistically grounded interpretations of TOEFL ITP scores can be integrated into an OBE-oriented ELT framework. Nevertheless, such alignment should be used as a complementary tool rather than a standalone measure, ensuring that assessment practices remain pedagogically meaningful and contextually appropriate.

ACKNOWLEDGEMENT

I would like to express gratitude to the anonymous reviewers for their insightful comments and

constructive suggestions. I also extend my appreciation to the editorial team for their efficient handling of the review process and their valuable support throughout the publication process.

REFERENCES

Alderson, J. C. (2005). *Diagnosing foreign language proficiency*. Continuum.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2020). *Standards for educational and psychological testing*. American Educational Research Association.

Ary, D., Jacobs, L. C., Irvine, C. K. S., & Walker, D. A. (2023). *Introduction to research in education* (11th ed.). Cengage Learning.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford University Press.

Biggs, J., & Tang, C. (2022). *Teaching for quality learning at university* (5th ed.). McGraw-Hill Education.

Brennan, R. L. (Ed.). (2006). *Educational measurement* (4th ed.). Praeger.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. Pearson Education.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2021). *Building a validity argument for language assessment*. Routledge. https://doi.org/10.4324/9781003164255

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

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

Council of Europe. (2020). *Common European framework of reference for languages: Learning, teaching, assessment—Companion volume*. Council of Europe Publishing.

Creswell, J. W., & Creswell, J. D. (2023). *Research design: Qualitative, quantitative, and mixed methods approaches* (6th ed.). SAGE Publications.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Prentice Hall.

Educational Testing Service. (2023). *TOEFL ITP® assessment series: Technical manual*. ETS.

Field, A. (2022). *Discovering statistics using IBM SPSS statistics* (6th ed.). SAGE Publications.

Fulcher, G. (2010). *Practical language testing*. Routledge.

Ginther, A., & Yan, X. (2018). TOEFL ITP® test performance and validity considerations. *ETS Research Report Series*, *2018*(1), 1–15. https://doi.org/10.1002/ets2.12185

Green, A. (2014). *Exploring language assessment and testing: Language in action*. Routledge.

Haertel, E. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Praeger.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.

Harden, R. M. (2019). Outcome-based education. In *Understanding medical education* (pp. 111–124). Wiley. https://doi.org/10.1002/9781119373780.ch8

Harsch, C. (2014). General language proficiency. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 457–472). Wiley.

Harsch, C., & Hartig, J. (2015). What are we aligning tests to? The construct of language proficiency in the CEFR. *Language Testing*, *32*(3), 321–340. https://doi.org/10.1177/0265532214567593

Hidri, S. (2020). *Language assessment in the Middle East and North Africa: Theory, practice, and policy*. Springer. https://doi.org/10.1007/978-3-030-53254-3

Huhta, A., Johnson, E., & Tulkki, H. (2023). CEFR alignment in higher education contexts. *Language Assessment Quarterly*, *20*(2), 120–138. https://doi.org/10.1080/15434303.2022.2150000

Islam, S. I. U., & Sood, S. (2023). Outcome-based education: A systematic review. *European Chemical Bulletin*, *12*(Special Issue 4), 1–10.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kane, M. T. (2021). Explicating validity. *Educational Measurement: Issues and Practice*, *40*(3), 3–11. https://doi.org/10.1111/emip.12401

Karantonis, A., & Sireci, S. G. (2022). Standard setting methods: A review. *Educational Measurement: Issues and Practice*, *41*(2), 45–57. https://doi.org/10.1111/emip.12477

Kolen, M. J., & Brennan, R. L. (2020). *Test equating, scaling, and linking: Methods and practices* (4th ed.). Springer. https://doi.org/10.1007/978-1-4939-0317-7

Kunnan, A. J. (2021). *Evaluating language assessments* (2nd ed.). Routledge.

Lane, S., Raymond, M. R., & Haladyna, T. M. (Eds.). (2016). *Handbook of test development* (2nd ed.). Routledge.

Leung, C., & Scarino, A. (2021). Reconceptualizing the nature of goals and outcomes in language/s education. *Modern Language Journal*, *105*(S1), 16–33. https://doi.org/10.1111/modl.12686

Little, D., & Figueras, N. (2022). The CEFR and language policy. *Language Teaching*, *55*(3), 1–15. https://doi.org/10.1017/S0261444821000273

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Peter Lang.

McNamara, T. (2000). *Language testing*. Oxford University Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.

Mislevy, R. J. (2018). Sociocognitive foundations of assessment. *Educational Measurement: Issues and Practice*, *37*(2), 10–20. https://doi.org/10.1111/emip.12186

North, B. (2014). *The CEFR in practice*. Cambridge University Press.

North, B., Figueras, N., Takala, S., & Verhelst, N. (2010). Linking language examinations to the CEFR: A manual. *Language Testing*, *27*(3), 323–340. https://doi.org/10.1177/0265532210373906

North, B., & Piccardo, E. (2023). Developing illustrative descriptors of aspects of mediation for the CEFR. *Language Teaching*, *56*(1), 78–95. https://doi.org/10.1017/S026144482100033X

O'Sullivan, B. (2011). *Language testing: Theory and practice*. Palgrave Macmillan.

Papageorgiou, S., Bridgeman, B., Tannenbaum, R. J., & Cho, Y. (2021). Linking TOEFL iBT® scores to the CEFR: Methodological considerations. *Language Testing*, *38*(3), 455–478. https://doi.org/10.1177/0265532211003077

Piccardo, E. (2020). The Common European Framework of Reference (CEFR) in language education: Past, present, and future. *Language Teaching*, *53*(4), 475–494. https://doi.org/10.1017/S0261444820000146

Powers, D. E., & Schedl, M. (2021). Linking TOEFL iBT® scores to the CEFR: A research synthesis. *ETS Research Report Series, 2021*(1), 1–23. https://doi.org/10.1002/ets2.12345

Rao, N. J. (2020). Outcome-based education: An overview. *Journal of Educational Research and Practice*, *10*(1), 1–8.

Read, J. (2000). *Assessing vocabulary*. Cambridge University Press.

Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2023). Outcome-based education in the digital era. *Educational Technology Research and Development*, *71*(2), 345–362. https://doi.org/10.1007/s11423-022-10123-4

Sireci, S. G. (2021). Standard setting and score interpretation. *Educational Measurement: Issues and Practice*, *40*(1), 25–32. https://doi.org/10.1111/emip.12390

Spady, W. G. (1994). *Outcome-based education: Critical issues and answers*. American Association of School Administrators.

Tannenbaum, R. J., & Wylie, E. C. (2008). Linking English-language test scores onto the CEFR. *ETS Research Report Series*, *2008*(1), 1–51. https://doi.org/10.1002/j.2333-8504.2008.tb02110.x

Taylor, L. (2014). *A report on the association of English language test scores with CEFR levels*. Cambridge English.

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

Wiliam, D. (2011). *Embedded formative assessment*. Solution Tree Press.

Xu, Z., Wang, Y., & Li, H. (2024). Statistical validation of CEFR alignment in large-scale language testing. *Language Testing*, *41*(2), 215–234. https://doi.org/10.1177/02655322231234567