EDUKASIA: Jurnal Pendidikan dan Pembelajaran
Vol. 7, 1 (January-June, 2026), pp. 203-218
ISSN: 2721-1150; EISSN: 2721-1169; DOI: 10.62775/edukasia.v7i1.1842

Psychometric Properties of the Classroom English Proficiency (CEP) Scale for English Language Teachers in Indonesia: A CFA and Rasch Model Analysis

Syahadah Albaqiyatul Karimah[1], Devie Yundianto[1], Muhammad Aqil Alaauddin[1]

[1] Universitas Nahdlatul Ulama Indonesia, Indonesia

ARTICLE INFO

Keywords:

Classroom English Proficiency Scale;
Confirmatory Factor Analysis;
English Teacher;
Psychometric Properties;
Rasch Model

Article history:

Received 2025-11-19
Revised 2026-01-08
Accepted 2026-02-22

ABSTRACT

This study aimed to adapt and validate the Classroom English Proficiency Scale (CEPS) for English language teachers in Indonesia using Confirmatory Factor Analysis (CFA) and the Rasch Rating Scale Model (RSM). The CFA confirmed that the four latent dimensions—Grammar, Pronunciation, Interaction, and Instruction—fit well within a second-order structure representing a single construct of classroom English proficiency. Rasch analysis further supported the unidimensionality assumption, with variance explained by measures exceeding the recommended criterion. The scale showed high reliability, and all items met model-fit expectations. The five-point Likert rating scale functioned effectively, with ordered Andrich thresholds and consistent category use. Differential Item Functioning (DIF) analysis indicated that 11 of 12 items were invariant across gender groups, except CEPS_2 ("I can use a wide range of English vocabulary"), which was slightly more difficult for male respondents. Overall, the Indonesian-adapted CEPS demonstrated strong validity, reliability, and fairness, confirming its suitability for assessing English proficiency in classroom contexts. The CEPS can serve as a reliable diagnostic and evaluative tool for English language teachers in teacher education and EFL assessment settings.

This is an open access article under the CC BY-NC-SA license.

Corresponding Author:
Syahadah Albaqiyatul Karimah
Universitas Nahdlatul Ulama Indonesia, Indonesia; syahalbakarimah@unusia.ac.id

1. INTRODUCTION

In today's era, English proficiency is no longer considered an additional skill but rather a necessity, particularly in the field of education. For teachers, English proficiency is not only about mastering grammar and vocabulary but also encompasses the ability to deliver lessons clearly, communicate with students, and manage classrooms effectively in English (Numonova, 2024); (Zega, 2025); (Sahnan & Daulay, 2025). Teachers with strong English competence are believed to foster more interactive learning

environments and motivate students to develop greater confidence in using English (Ali et al., 2020); (Al-Barakat et al., 2025); (Zhang, 2025).

Language proficiency is regarded as one of the core competencies that English teachers must possess (Pham, 2018); (Ismailov et al., 2025). However, general English proficiency alone does not guarantee teachers' readiness for classroom instruction, as teaching requires context-specific language use that supports pedagogical functions (Chambless, 2012); (Pham, 2018). Teachers with insufficient classroom-related language skills often rely heavily on textbooks and scripted materials, limiting spontaneous interaction and authentic communication with learners (Medgyes P, 2001); (Nyström, 2025); (Nguyen, 2025); (Aramaki, 2025).

In Indonesia, Classroom English Proficiency (CEP) has gained scholarly attention in recent years due to its pivotal role in enhancing the effectiveness of English instruction. CEP refers to the level of English proficiency required to teach effectively in classroom contexts, encompassing essential classroom-related skills such as giving instructions, managing interaction, providing feedback, and scaffolding students' understanding (Freeman, 2017); (Wang, 2021); (Matsumura & Hinoki, 2024); (Walsh, 2011); (Richards, 2017). Grounded in the communicative competence framework (Canale & Swain, 1980), CEP reflects the integration of linguistic, sociolinguistic, discourse, and strategic competences situated within teaching practices. Empirical studies consistently demonstrate that higher levels of teachers' English proficiency are associated with more effective classroom management, clearer lesson delivery, and increased student engagement (Goh & Burns, 2012); (Walsh, 2011). Nevertheless, teaching effectiveness is also influenced by complementary factors such as teacher confidence, professional experience, and institutional support (Narzillayevna, 2024), underscoring the need for valid and reliable measurement of classroom-specific language proficiency.

Several instruments have been developed to assess CEP, including the English Language Teaching Confidence Scale (ELT-CS) (Chacón, 2005) and the Teacher Oral Proficiency in English Scale (TOPE) (DE JONG et al., 2013). These instruments primarily focus on oral proficiency, fluency, and teachers' confidence in using English for instructional purposes. However, prior research also emphasizes that CEP development is shaped by contextual factors such as teacher education, classroom environment, and pedagogical innovation, and that native-speaker status alone does not ensure effective teaching competence (Kamhi-Stein, 2016); (Waddington, 2022); (Selvi et al., 2024).

Despite the growing recognition of CEP, no measurement scale has been specifically adapted and psychometrically validated for the Indonesian educational context. Recent Indonesian studies highlight the importance of localized validation to ensure cultural relevance and measurement accuracy (Karimah et al., 2025). This gap indicates the need for a rigorous psychometric evaluation of CEP instruments that aligns with both local teaching practices and contemporary measurement standards.

Addressing this gap, the present study aims to adapt and psychometrically validate the Classroom English Proficiency Scale (CEPS) for English language teachers in Indonesia using an integrated Confirmatory Factor Analysis (CFA) and Rasch modeling framework. Specifically, this study investigates whether the CEP scale demonstrates adequate validity, reliability, unidimensionality, and appropriate item difficulty to support teacher assessment and professional development in the Indonesian context.

2. METHODS

2.1 Participants

This study employed a quantitative research method with a scale development approach to examine English language teachers in Indonesia. A total of 202 participants voluntarily took part in this study. Participants were recruited using a convenience sampling technique. All participants were active English teachers from both formal and non-formal educational institutions, including public and

private schools, and represented various educational levels such as primary, secondary, and tertiary education. The sample consisted of 165 females (81.7%) and 37 males (18.3%), with ages ranging from 18 to 52 years (M = 25.80, SD = 5.4). Prior to participation, all respondents provided informed consent, and their anonymity and confidentiality were strictly maintained.

2.2 Research Instrument

The instrument used in this study was the Classroom English Proficiency Scale (CEPS), originally developed by (Wang, 2021), which was adapted and modified into an Indonesian version. The Classroom English Proficiency Scale (CEPS) comprises 12 items that represent four key dimensions of teachers' English proficiency in the classroom.

The first dimension, Grammatical and Lexical Accuracy and Range (3 items), measures teachers' ability to use correct grammar and a rich vocabulary in classroom communication. The second dimension, Pronunciation, Stress, and Intonation (3 items), assesses the clarity of pronunciation, accuracy of word stress, and appropriateness of intonation patterns in spoken English. The third dimension, The Language of Interaction (3 items), evaluates teachers' ability to use English effectively for classroom interaction, such as giving feedback, managing discussions, and responding to students. The final dimension, The Language of Instruction (3 items), measures teachers' competence in understanding and using English as a medium of instruction to facilitate learning and deliver lesson content clearly and effectively.

All items were presented in the form of statements rated on a five-point Likert scale, ranging from 1 (Strongly Disagree) to 5 (Strongly Agree). A higher score on each dimension indicates a greater perceived level of English proficiency. The back-translation technique was applied to ensure equivalence between the original and the Indonesian version. The development of the CEPS items was based on an extensive review of the literature on language proficiency assessment and was qualitatively validated through expert judgment by two specialists in the field of English language teaching and psychology to ensure content validity.

2.3 Data Collection Procedures

The data collection process was conducted through both online and offline methods. The online survey was distributed via Google Forms, while the offline data were collected by directly visiting schools and educational institutions. On the initial page of the questionnaire, the researchers provided a comprehensive explanation of the study's objectives, assurances of data confidentiality, and the voluntary nature of participation. Participants who agreed to participate proceeded to complete the questionnaire. The data collection process lasted approximately 10 weeks. Upon completion of data collection, all responses were prepared and organized for statistical analysis.

2.4 Data Analysis Procedures

Data were analyzed using the Rating Scale Model (RSM) developed by (Andrich, 1978). RSM is a measurement model for polytomous data (data with two or more ordinal categories). The model provides estimates of person locations, item difficulties, and overall thresholds (fixed across items), while allowing item difficulties to vary. To address the need for a valid and reliable instrument in measuring Classroom English Proficiency (CEP) among teachers in Indonesia, this study adopts a quantitative approach with a cross-sectional design. The study aims to analyze the appropriateness and accuracy of the CEP instrument as a measure of English proficiency in classroom teaching, while also validating and further developing the scale to align with the characteristics of Indonesian teachers.

2.5 Confirmatory Factor Analysis (CFA)

Data analysis was performed using JASP statistical software version 0.95.3.0. Several statistical procedures were employed to examine the psychometric properties of the Classroom English Proficiency Scale (CEPS). First, descriptive statistical analyses were conducted to identify participants' demographic characteristics and to summarize the distribution of scores for each item. Second, to investigate the internal structure of the scale, a Confirmatory Factor Analysis (CFA) was carried out. CFA was selected because the present study adapted the CEPS instrument based on the theoretical framework of the original measure, which specifies the presence of four latent factors. The proposed four-factor model was tested to determine its fit with the empirical data.

Model fit was evaluated using several goodness-of-fit indices, including Chi-square (χ^2), Comparative Fit Index (CFI), Tucker–Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMR). The model was considered to have an acceptable fit if it met the recommended thresholds of CFI and TLI > 0.90, RMSEA < 0.08, and SRMR < 0.08 (Hu & Bentler, 1999; Wang & Wang, 2019). Third, reliability analysis was conducted to evaluate the internal consistency of each CEPS subscale. Reliability coefficients were calculated using Cronbach's Alpha and McDonald's Omega, with values greater than 0.70 indicating acceptable reliability (Bentler, 2017); (Bodoff, 2008).

2.6 Rasch Model Analysis

The study also hypothesized that a unidimensional model would fit the data according to the Rasch Model requirements. Data were analyzed using Winsteps software version 3.73. The Rasch Model was employed for its ability to convert ordinal response scores into linear interval measures (logits) for both person ability and item difficulty (Bond, 2015). Specifically, the Rating Scale Model (RSM) was applied, as all CEPS items share the same response format (Chong et al., 2021). In this study, the unconditional maximum likelihood estimator was used for RSM, while Maximum Likelihood estimation was used in the CFA procedure.

2.7 Rating Scale Functioning

The initial stage of Rasch analysis involved examining the functioning of the rating scale categories to ensure that the Likert response options operated as intended. Evaluation criteria included: (a) each response category having a minimum frequency of 10; (b) mean-square (Outfit MNSQ) values for each category being less than 2.0; and (c) threshold values (step calibrations) increasing monotonically, indicating that higher response categories consistently represented higher proficiency levels (Linacre, 2002).

2.8 Unidimensionality

A core assumption of the Rasch Model is unidimensionality, meaning the instrument measures a single dominant construct (Medvedev & Krägeloh, 2025); (Santoso et al., 2025).This assumption was evaluated using the Principal Component Analysis of Residuals (PCAR). The scale was considered unidimensional if: (a) the Raw Variance Explained by Measures (RVEM) was at least 40% (Holster & Lake, 2016), and (b) the Unexplained Variance in the First Contrast (UVIC) had an eigenvalue not exceeding 15% (Fan & Bond, 2019).

2.9 Item and Person Fit

Item and person fit were assessed using Mean Square (MNSQ) statistics, which include Infit and Outfit indices (Rahman, 2023). Infit MNSQ is sensitive to unexpected response patterns on items whose

difficulty levels are close to the respondent's ability (Megbele et al., 2023), whereas Outfit MNSQ is more sensitive to unexpected responses on items that are much easier or harder. Items were considered to fit the model if both Infit and Outfit MNSQ values ranged between 0.6 and 1.4, and point–measure correlations were positive (Bond et al., 2020).

2.10 Reliability and Separation

Rasch reliability was estimated separately for items and persons. Item reliability indicates how well items define a stable hierarchy of difficulty, while person reliability shows how effectively the instrument distinguishes among participants' ability levels. Additionally, separation indices were examined to determine how many statistically distinct strata could be identified for both items and persons (Bond et al., 2020).

2.11 Item–Person Wright Map

The final stage of analysis involved constructing an Item–Person Wright Map to visualize the distribution of person abilities and item difficulties along the same logit scale. This map is useful for evaluating targeting, that is, whether the item difficulty levels appropriately match the range of participant abilities measured by the scale (Bond et al., 2020).

2.12 Differential Item Functioning (DIF)

Differential Item Functioning (DIF) occurs when an item exhibits different interpretations across groups, resulting in one group being advantaged or disadvantaged (Hambleton & Jones, 1994). Rasch modeling emphasizes the importance of items being free from bias (Christensen et al., 2013). DIF was considered substantial when the DIF contrast reached 0.5 logits. Furthermore, items were flagged for potential elimination if the Rasch–Welch t-test exceeded ±1.96 and was statistically significant ($p < 0.05$) (Linacre, 2018).

To enhance clarity and transparency, Table 1 summarizes the key methodological procedures and analytical steps employed in this study. The table provides an overview of the sequential stages of instrument adaptation, data collection, and psychometric evaluation using Confirmatory Factor Analysis (CFA) and Rasch modeling. This summary is intended to facilitate reproducibility and to clearly illustrate how each analytical procedure aligns with the study objectives and validation framework.

Table 1 Summary of Methodological Procedures and Analytical Steps

Stage	Procedure	Purpose	Method / Software
Instrument adaptation	Translation and back-translation	Linguistic and cultural equivalence	Expert judgment
Content validation	Expert review	Content validity	Qualitative evaluation
Data collection	Online and offline survey	Data acquisition	Google Forms & field visits
Construct validation	Confirmatory Factor Analysis (CFA)	Factor structure and construct validity	JASP
Reliability analysis	Internal consistency	Scale reliability	Cronbach's Alpha, McDonald's Omega
Item-level analysis	Rasch Rating Scale Model (RSM)	Item difficulty and person ability	Winsteps
Rating scale evaluation	Category functioning	Response scale effectiveness	Rasch analysis

Unidimensionality testing	PCAR	Dimensionality assessment	Rasch analysis
Fairness analysis	Differential Item Functioning (DIF)	Gender invariance	Rasch–Welch t-test

3. FINDINGS AND DISCUSSION

The data analysis begins with descriptive statistics and tests of normality assumptions. Subsequently, the results of a second-order Confirmatory Factor Analysis (CFA) are presented to provide preliminary evidence for the construct validity of the scale. The main section of this chapter then reports the results of an in-depth psychometric analysis using the Rasch Rating Scale Model, aimed at evaluating the quality and measurement precision of the Classroom English Proficiency Scale (CEPS).

3.1 Descriptive Analysis

Descriptive analysis was conducted on the 12 CEPS items collected from 202 participants. The normality test results indicated that the data were normally distributed. The skewness values for the 12 items ranged from –1.461 (CEPS 12) to –0.422 (CEPS 1), while the kurtosis values ranged from –0.312 (CEPS 1) to 2.496 (CEPS 12). All skewness and kurtosis values were below the absolute cutoff points recommended for multivariate analyses (|2.0| for skewness and |7.0| for kurtosis). Therefore, the dataset met the assumptions required for further multivariate analyses, including CFA.

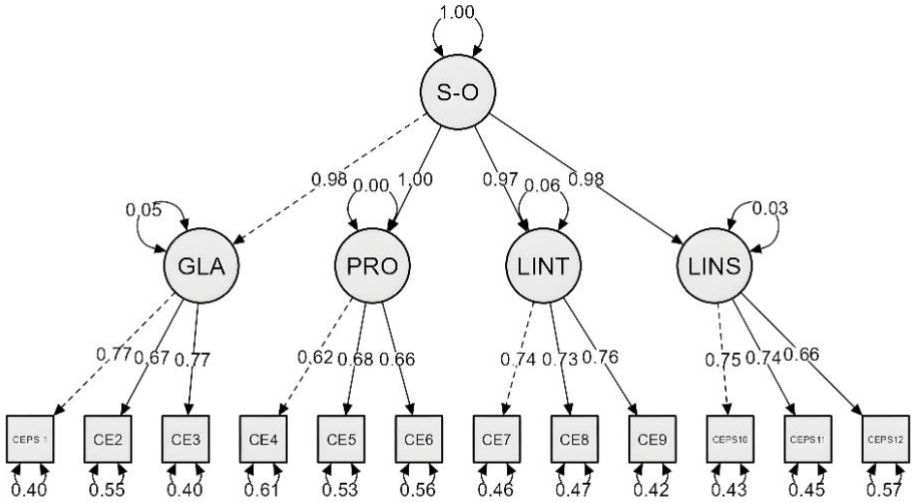
3.2 Confirmatory Factor Analysis (CFA)

As an initial step in testing construct validity, a second-order CFA was performed to determine whether the proposed four-factor model, unified by a higher-order factor, adequately fit the empirical data.

The model fit indices demonstrated an overall good fit between the hypothesized model and the observed data (see Figure 1). Although the Chi-square value was significant ($\chi^2(50) = 97.432$, $p < 0.001$)—a common occurrence in larger samples—the other fit indices supported a robust model. The Comparative Fit Index (CFI) was 0.960, and the Tucker–Lewis Index (TLI) was 0.947, both exceeding the recommended threshold of 0.90. Furthermore, the error indices indicated acceptable model fit: the Root Mean Square Error of Approximation (RMSEA) was 0.069 (90% CI [0.048, 0.089]), and the Standardized Root Mean Square Residual (SRMR) was 0.043. An SRMR value below 0.08 suggests a very good model fit.

The factor loading analysis revealed that all 12 items were significant indicators of their respective first-order latent factors ($p < 0.001$), with standardized estimates ranging from 0.623 (CEPS 4) to 0.774 (CEPS 1 and CEPS 3). Moreover, the four first-order factors were also found to significantly load onto a second-order construct, with standardized loadings ranging from 0.969 to 1.000, providing strong evidence of hierarchical construct validity.

From the perspective of Classical Test Theory (CTT), the overall reliability of the CEPS instrument was high, with a Cronbach's alpha (α) coefficient of 0.920. The internal consistency coefficients for each subscale were as follows: Grammatical and Lexical Accuracy and Range ($\alpha = 0.782$), Pronunciation, Stress, and Intonation ($\alpha = 0.682$), Language of Interaction ($\alpha = 0.785$), and Language of Instruction ($\alpha = 0.760$). These results demonstrate that the CEPS possesses strong internal reliability and consistent measurement across its four dimensions.



GLA = Grammatical and lexical accuracy and range; PRO = Pronunciation, stress, and intonation; LINT = The language of interaction; LINS = The language of instruction

Figure 1 Four-Dimensional Second-Order Factor Measurement Model

3.3 Rasch Model Analysis

To obtain a more in-depth psychometric evaluation at the interval level, the analysis was further conducted using the Rasch Rating Scale Model (RSM), implemented with Winsteps version 3.73.

Unidimensionality

The fundamental assumption of the Rasch model is unidimensionality, which means that the instrument measures a single dominant construct. This assumption was tested using Principal Component Analysis of Residuals (PCAR). The analysis results confirmed that the CEPS instrument strongly meets the unidimensionality assumption. The Raw Variance Explained by Measures (RVEM) was 49.2%, exceeding the recommended minimum threshold of 40%, indicating that one primary construct accounts for the majority of variance in the data.

Rasch Reliability and Separation

The summary statistics indicated excellent levels of reliability and separation. For the participants (N = 202), the Person Reliability value was 0.85, reflecting a high level of response consistency. The Person Separation Index (PSI) was 2.38, suggesting that the instrument can distinguish participants into at least two to three statistically distinct ability strata.

For the items, the Item Reliability was 0.93, indicating that the hierarchy of item difficulty levels was highly stable and replicable. The Item Separation Index of 3.76 shows that the items can effectively differentiate three to four distinct difficulty levels. As confirmation, the classical reliability estimate (Cronbach's Alpha/KR-20) calculated by Winsteps was 0.92, which aligns with the Rasch reliability indices.

Item Fit and Item-Person Map

The item fit analysis was conducted to identify any items that functioned anomalously within the Rasch model. The criteria used were Infit and Outfit Mean Square (MNSQ) values within the ideal range of 0.6 to 1.4, and Standardized Fit Statistics (ZSTD) within the ideal range of –2.0 to +2.0. These criteria ensured that all items contributed meaningfully to measuring the intended construct.

Table 2 Item Fit Statistic

Item	Original Item	Item Wording	Measure	MNSQ		PTMEA
				Infit	Outfit	
CEPS_1	I can lecture with correct English grammatical structures.	*Saya dapat memberikan kuliah dengan struktur tata bahasa Inggris yang benar*	0.3	0.77	0.79	0.77
CEPS_2	I can use a broad range of English vocabulary.	*Saya dapat menggunakan berbagai kosa kata bahasa Inggris yang luas*	0.04	1.07	1.1	0.69
CEPS_3	I can use accurate words to express ideas.	*Saya dapat menggunakan kata-kata yang tepat untuk mengungkapkan ide*	0.2	0.8	0.79	0.75
CEPS_4	I can speak English clearly with no systematic errors in pronunciation.	*Saya dapat berbicara bahasa Inggris dengan jelas tanpa kesalahan sistematis dalam pengucapan*	1.44	1.38	1.39	0.68
CEPS_5	I know how to stress content words in pronunciation.	*Saya tahu cara memberi tekanan pada kata-kata bermakna (content words) dalam pengucapan*	0.46	1.09	1.12	0.7
CEPS_6	I can use intonation naturally to convey meaning.	*Saya dapat menggunakan intonasi secara alami untuk menyampaikan makna*	-0.15	1.11	1.05	0.68
CEPS_7	I can use appropriate English to ask questions or to provide clues and hints.	*Saya dapat menggunakan bahasa Inggris yang sesuai untuk mengajukan pertanyaan atau memberikan petunjuk dan petanda*	-0.29	0.95	0.95	0.7
CEPS_8	I can use appropriate English to respond to students' questions, such as seeking clarification, giving confirmation, and asking for repetition.	*Saya dapat menggunakan bahasa Inggris yang sesuai untuk merespons pertanyaan siswa, seperti meminta klarifikasi, memberikan konfirmasi, dan meminta pengulangan*	-0.4	0.89	0.86	0.71
CEPS_9	I can give feedback skillfully in English, such as acknowledging, evaluating, and commenting on	*Saya dapat memberikan umpan balik dengan terampil dalam bahasa Inggris, seperti mengakui, mengevaluasi, dan mengomentari respons siswa*	-0.15	0.83	0.78	0.72

	students' responses.					
CEPS_10	I can explain concepts, terms, or lesson content in clear English.	*Saya dapat menjelaskan konsep, istilah, atau materi pelajaran dengan bahasa Inggris yang jelas*	-0.1	0.89	0.84	0.73
CEPS_11	I can give clear instructions in English when conducting activities, giving homework, and managing the classroom.	*Saya dapat memberikan instruksi yang jelas dalam bahasa Inggris saat melaksanakan aktivitas, memberikan pekerjaan rumah, dan mengelola kelas*	-0.23	1	0.95	0.71
CEPS_12	I can use appropriate English signals (e.g., first, second, next) to indicate stages of a lesson.	*Saya dapat menggunakan penanda bahasa Inggris yang tepat (misalnya: first, second, next) untuk menunjukkan tahapan dalam suatu pelajaran*	-1.11	1.21	1.23	0.61

Based on the results presented in Table 2, all CEPS items demonstrated a good fit with the Rasch model. Overall, none of the items in the Indonesian-adapted version of the CEPS instrument were eliminated, indicating strong internal consistency and item validity. Among the items, CEPS_4 ("I can speak English clearly without systematic pronunciation errors") was identified as the most difficult item, meaning it was the least frequently experienced or achieved by participants. Conversely, CEPS_12 ("I can use appropriate English discourse markers such as first, second, and next to indicate lesson stages") was found to be the easiest item, meaning it was the most experienced by participants. The hierarchy of item difficulty levels is illustrated further in Figure 2.

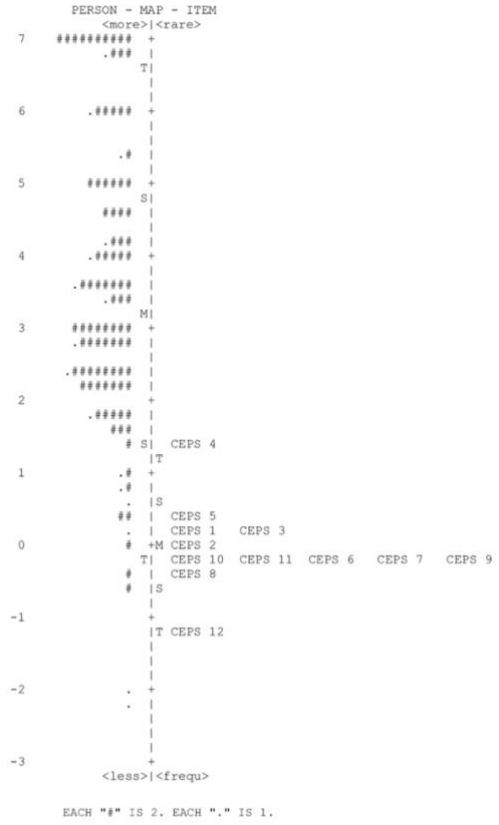


Figure 2 Item Person Map CEPS

Evaluation of Rating Scale Function

The functioning of the five-point Likert rating scale was examined using the Rating Scale Model (RSM). The results indicated that all response categories were utilized by participants, although the lowest category ("Strongly Disagree") showed a relatively low frequency of endorsement.

The rating scale met the monotonicity assumption, as both observed averages and Andrich threshold estimates increased progressively across response categories, indicating orderly category functioning. Category fit statistics showed acceptable performance for categories 2 to 5 (Infit and Outfit MNSQ between 0.92 and 1.06). Although category 1 exhibited misfit due to its low frequency, the overall rating scale functioned effectively and consistently for measuring Classroom English Proficiency.

Measurement Invariance

Differential Item Functioning (DIF) analysis was conducted to examine whether the items functioned equivalently across different subgroups. In this study, DIF was tested based on gender (male vs. female). The results indicated that 11 out of 12 items functioned fairly and showed no significant DIF. However, one item—CEPS 2 ("I can use a wide range of English vocabulary")—demonstrated statistically significant DIF.

The Welch t-test indicated a significant difference ($t = 2.94$, $p = .0045$), which was further confirmed by the Mantel–Haenszel test ($\chi^2 = 11.649$, $p = .0006$). The item was found to be more difficult for the male group (DIF measure = 0.87) than for the female group (DIF measure = –0.17). The DIF contrast of 1.04 logits indicates a substantial difference in item difficulty between the two groups.

3.4 Discussion

This study aimed to adapt and evaluate the psychometric properties of the Classroom English Proficiency Scale (CEPS) using two complementary analytical frameworks: Confirmatory Factor Analysis (CFA) and the Rasch Model. Overall, the findings indicate that CEPS is a psychometrically robust and theoretically sound instrument for assessing classroom-specific English proficiency among English teachers in Indonesia.

The strongest evidence emerges from the construct validity results. The second-order CFA demonstrated excellent model fit, with all major goodness-of-fit indices meeting recommended thresholds (CFI = 0.960; TLI = 0.947; RMSEA = 0.069, 90% CI [0.048, 0.089]; SRMR = 0.043). These findings confirm that the four theoretically grounded dimensions—Grammatical and Lexical Accuracy and Range, Pronunciation, Stress, and Intonation, Language of Interaction, and Language of Instruction—are coherently represented by a higher-order construct of Classroom English Proficiency. This hierarchical structure supports the conceptualization of CEPS as multidimensional at the subscale level while remaining unidimensional at the global construct level.

This interpretation is further strengthened by the Rasch analysis, which confirmed the unidimensionality assumption through Principal Component Analysis of Residuals. The Raw Variance Explained by Measures (49.2%) exceeded the recommended threshold of 40%, indicating that a single dominant latent trait underlies item responses. Together, the CFA and Rasch results provide converging evidence that CEPS can validly yield both subscale scores and an overall proficiency score, enhancing its flexibility for research and professional development purposes.

In terms of measurement precision, CEPS demonstrated strong reliability across analytical frameworks. High person reliability (0.85) indicates that the scale can effectively distinguish teachers across different levels of classroom English proficiency, while high item reliability (0.93) suggests that the item difficulty hierarchy is stable and replicable. These results confirm that CEPS possesses sufficient measurement sensitivity for both diagnostic and evaluative applications.

A noteworthy finding concerns the presence of Differential Item Functioning (DIF) in CEPS_2 ("I can use a wide range of English vocabulary") across gender groups. The analysis revealed that this item was more difficult for male teachers than for female teachers with equivalent levels of overall proficiency. From a practical perspective, this finding has important implications for scale use and development. Rather than warranting immediate item deletion, the observed DIF suggests a need for careful item review. The item may reflect differences in self-perception of lexical breadth rather than actual classroom language competence, potentially influenced by gender-related response tendencies. Therefore, revision rather than removal appears to be the most appropriate course of action. Future iterations of CEPS may benefit from rephrasing the item to include more behaviorally anchored descriptors or contextualized classroom examples, thereby reducing subjective interpretation while preserving its conceptual relevance.

Despite the strong psychometric evidence, several limitations of the present study should be explicitly acknowledged. First, the use of convenience sampling restricts the generalizability of the findings to the broader population of Indonesian English teachers. Second, the reliance on self-report data introduces the possibility of response bias, as participants' ratings may not fully reflect actual classroom language performance. These limitations highlight the need for future validation studies employing probabilistic sampling techniques and incorporating external criteria, such as classroom observations or performance-based assessments. Longitudinal studies are also recommended to examine the stability and sensitivity of CEPS across time and instructional contexts.

4. CONCLUSION

This study aimed to adapt, validate, and evaluate the psychometric properties of the Classroom English Proficiency Scale (CEPS) for English language teachers in Indonesia using a combined analytical framework of Confirmatory Factor Analysis (CFA) and the Rasch Rating Scale Model (RSM). Overall, the findings provide strong empirical evidence that the Indonesian-adapted CEPS demonstrates satisfactory construct validity, reliability, and measurement precision.

Based on the stated research objectives, the main conclusions of this study can be summarized as follows:

1. The CFA results confirm that the proposed four-factor structure—Grammatical and Lexical Accuracy and Range, Pronunciation, Stress, and Intonation, Language of Interaction, and Language of Instruction—fits the empirical data well and is coherently represented by a higher-order construct of Classroom English Proficiency.
2. Rasch analysis provides further support for the unidimensionality of the CEPS, with the scale measuring a single dominant latent trait while maintaining theoretically meaningful subdimensions.
3. The instrument demonstrates strong reliability and separation indices for both persons (0.85) and items (0.93), indicating its effectiveness in distinguishing different levels of classroom English proficiency among teachers.
4. All items showed acceptable fit to the Rasch model, and the Wright Map indicated that item difficulty levels were well aligned with participants' ability levels.
5. Although one item (CEPS_2) exhibited gender-related Differential Item Functioning (DIF), its impact on the overall psychometric integrity of the scale was limited, supporting the general fairness of the instrument across gender groups.

Taken together, these findings indicate that the Indonesian version of CEPS is a valid, reliable, and practically useful instrument for assessing English teachers' classroom language proficiency. The scale can serve both diagnostic and evaluative functions in teacher education programs, professional development initiatives, and institutional language assessment practices. Future research is encouraged to replicate the present findings using larger and more diverse samples to further strengthen generalizability.

In addition, longitudinal studies are recommended to examine the sensitivity of CEPS in capturing changes in teachers' classroom English proficiency over time, as well as to explore the integration of CEPS with observational or performance-based assessment methods.

Acknowledgments: This research was funded by BIMA Kemendikbudristek (Ministry of Education, Culture, Research, and Technology). The researcher would like to express our deepest gratitude to the Kemendikbudristek for the support and funding provided through the BIMA research grant. Thank you to the Universitas Nahdlatul Ulama Indonesia for supporting the research publication.

REFERENCES

Al-Barakat, A. A., Al-Hassan, O. M., AlAli, R. M., Bataineh, R. F., Aboud, Y. Z., & Ibrahim, N. A. (2025). Shaping Young Minds: How Teachers Foster Social Interaction, Psychological Security and Motivational Support in the Primary Language Classroom. *International Journal of Learning, Teaching and Educational Research*, *24*(1), 359–378. https://doi.org/10.26803/ijlter.24.1.18

Ali, Z., Masroor Farzana, & Khan, T. (2020). Creating Positive Classroom Environment For Learners' Motivation Towards Communicative Competence In The English Language. *Journal of the Research Society of Pakistan* , *57*(1), 317–328.

Andrich, D. (1978). Application of a psychometric model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, *2*(4), 581–594. https://doi.org/https://doi.org/10.1177/014662167800200413

Aramaki, T. (2025). Task-supported and task-based language teaching and their effects on task motivation. *The Journal of Educational Research*, 1–17. https://doi.org/10.1080/00220671.2025.2548577

Bentler, P. M. (2017). Specificity-enhanced reliability coefficients. *Psychological Methods*, *22*(3), 527–540. https://doi.org/10.1037/met0000092

Bodoff, D. (2008). Test theory for evaluating reliability of IR test collections. *Information Processing & Management*, *44*(3), 1117–1145. https://doi.org/10.1016/j.ipm.2007.11.006

Bond, T. (2015). *Applying the Rasch Model*. Routledge. https://doi.org/10.4324/9781315814698

Bond, T., Yan, Z., & Heene, M. (2020). *Applying the Rasch Model*. Routledge. https://doi.org/10.4324/9780429030499

Canale, M., & Swain, M. (1980). THEORETICAL BASES OF COMMUICATIVE APPROACHES TO SECOND LANGUAGE TEACHING AND TESTING. *Journal: APPLIED LINGUISTICS*, *1*, 1–47.

Chacón, C. T. (2005). Teachers' perceived efficacy among English as a foreign language teachers in middle schools in Venezuela. *Teaching and Teacher Education*, *21*(3), 257–272. https://doi.org/10.1016/j.tate.2005.01.001

Chambless, K. S. (2012). Teachers' Oral Proficiency in the Target Language: Research on Its Role in Language Teaching and Learning. *Foreign Language Annals*, *45*(s1). https://doi.org/10.1111/j.1944-9720.2012.01183.x

Chong, J., Mokshein, S. E., & Mustapha, R. (2021). Applying the Rasch Rating Scale Model (RSM) to investigate the rating scales function in survey research instrument. *Jurnal Cakrawala Pendidikan*, *41*(1), 97–111. https://doi.org/10.21831/cp.v41i1.39130

Christensen, K. B., Kreiner, S., & Mesbah, M. (2013). *Rasch Models in Health*. NJ: John Wiley & Sons, Inc.

DE JONG, N. H., STEINEL, M. P., FLORIJN, A., SCHOONEN, R., & HULSTIJN, J. H. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics*, *34*(5), 893–916. https://doi.org/10.1017/S0142716412000069

Fan, J., & Bond, T. (2019). Applying Rasch measurement in language assessment. In *Quantitative Data Analysis for Language Assessment Volume I* (pp. 83–102). Routledge. https://doi.org/10.4324/9781315187815-5

Freeman, D. (2017). The Case for Teachers' Classroom English Proficiency. *RELC Journal*, *48*(1), 31–52. https://doi.org/10.1177/0033688217691073

Goh, C. C. M., & Burns, A. (2012). *Teaching speaking: A holistic approach*. Cambridge University Press.

Hambleton, R. K., & Jones, R. W. (1994). Comparison of Empirical and Judgmental Procedures for Detecting Differential Item Functioning. *Educational Research Quarterly*, *18*(1), 21–36.

Holster, T. A., & Lake, J. (2016). Guessing and the Rasch Model. *Language Assessment Quarterly*, *13*(2), 124–141. https://doi.org/10.1080/15434303.2016.1160096

Ismailov, M., Chiu, T. K., Aizawa, I., Yamamoto, Y., Djalilova, N., & Moorhouse, B. L. (2025). Essential Lecturer Competencies in English Medium Instruction: A Study Across Student Proficiency Levels. *RELC Journal*. https://doi.org/10.1177/00336882241312427

Kamhi-Stein, L. D. (2016). The non-native English speaker teachers in TESOL movement. *ELT Journal*, *70*(2), 180–189. https://doi.org/10.1093/elt/ccv076

Karimah, S. A., Yundianto, D., Idris, M. M., & Zulha, H. F. (2025). PSYCHOMETRIC PROPERTIES OF SELF-EFFICACY SCALE FOR ENGLISH LANGUAGE LEARNERS IN INDONESIA. *English Review: Journal of English Education*, *13*(2), 517–528. https://doi.org/10.25134/erjee.v13i2.10954

Linacre, J. M. (2002). Understanding Rasch measurement: Optimizing Rating Scale Category Effectiveness. *Journal of Applied Measurement*, *3*(1), 85–106.

Linacre, J. M. (2018). *DIF-DPF-bias-interactions concepts. Winsteps Help for Rasch Analysis*. Https://Www.Winsteps.Com/Winman/Difconcepts.Htm.

Matsumura, S., & Hinoki, Y. (2024). Empowering Non-Specialist English Teachers: Self-Efficacy Enhancement Through Classroom English Proficiency and Collaborative Support. *Education Sciences*, *15*(1), 24. https://doi.org/10.3390/educsci15010024

Medgyes P. (2001). When the teacher is a non-native speaker. In: Celcie-Murcia M (ed.) English as a Second or Foreign Language. *Teaching Boston, MA: Heinle & Heinle*, 415–427.

Medvedev, O. N., & Krägeloh, C. U. (2025). Rasch Measurement Model. In *Handbook of Assessment in Mindfulness Research* (pp. 131–147). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-47219-0_4

Megbele, A. M., Odili, J. N., & Osadebe, P. U. (2023). ASSESSMENT OF THE PSYCHOMETRIC PROPERTIES OF ATTITUDE TOWARDS ASSESSMENT TEST (ATAT). *European Journal of Open Education and E-Learning Studies*, *8*(3). https://doi.org/10.46827/ejoe.v8i3.5054

Narzillayevna, A. S. (2024). THE IMPACT OF SELF-CONFIDENCE ON THE ACQUISITION OF ENGLISH LANGUAGE SKILLS: ASYSTEMATIC REVIEW. *Eurasian Journal of Entrepreneurship and Pedagogy*, *2*(1), 11–16.

Nguyen, T. A. (2025). Addressing English-speaking anxiety in Vietnamese initial teacher education: beyond coping strategies. *Asian. J. Second. Foreign. Lang. Educ*, *10*(24).

Numonova, M. (2024). COMMUNICATIVE METHODS OF TEACHING ENGLISH VOCABULARY AND GRAMMAR IN CONTEXT. *QO'QON UNIVERSITETI XABARNOMASI*, *13*, 327–330. https://doi.org/10.54613/ku.v13i.1090

Nyström, F. (2025). Opportunities for speaking: the if, how and when of students' spoken target language in the second foreign language classroom. *Moderna Språk*, *119*(2), 194–215. https://doi.org/10.58221/mosp.v119i2.24076

Pham, T. H. N. (2018). General English Proficiency or English for Teaching? The Preferences of In-service Teachers. *RELC Journal*, *49*(3), 339–352. https://doi.org/10.1177/0033688217691446

Rahman, Y. A. (2023). Person and Item Validity and Reliability in Essay Writing Using Rasch Model. *Konstruktivisme : Jurnal Pendidikan Dan Pembelajaran*, *15*(1), 41–55. https://doi.org/10.35457/konstruk.v15i1.2618

Richards, J. C. (2017). *Teacher language proficiency and classroom communication*. Cambridge University Press.

Sahnan, B., & Daulay, S. H. (2025). Developing Students' Vocabulary by Using Build-A-Sentence: Teachers' Perspective. *Scope : Journal of English Language Teaching*, *9*(2), 692. https://doi.org/10.30998/scope.v9i2.22547

Santoso, A., Afendi, F. M., Pardede, T., Retnawati, H., Rafi, I., Apino, E., & Rosyada, M. N. (2025). Deep-Rasch as an Alternative to Rasch Modeling under Assumption Violations and Small Sample Sizes. *CAUCHY: Jurnal Matematika Murni Dan Aplikasi*, *10*(2), 970–985. https://doi.org/10.18860/cauchy.v10i2.36276

Selvi, A. F., Yazan, B., & Mahboob, A. (2024). Research on "native" and "non-native" English-speaking teachers: Past developments, current status, and future directions. *Language Teaching*, *57*(1), 1–41. https://doi.org/10.1017/S0261444823000137

Waddington, J. (2022). Rethinking the 'ideal native speaker' teacher in early childhood education. *Language, Culture and Curriculum*, *35*(1), 1–17. https://doi.org/10.1080/07908318.2021.1898630

Walsh, S. (2011). *Exploring Classroom Discourse*. Routledge. https://doi.org/10.4324/9780203827826

Wang, C. (2021). The Relationship Between Teachers' Classroom English Proficiency and Their Teaching Self-Efficacy in an English Medium Instruction Context. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.611743

Zega, Y. S. (2025). Effective strategies for enhancing English speaking competence among learners in English education study programs. *Journal of Education, Social & Communication Studies*, *2*(2), 88–100. https://doi.org/10.71028/jescs.v2i2.27

Zhang, Q. (2025). The role of EFL teacher immediacy and teacher-student rapport in boosting motivation to learn and academic mindsets in online education. *Learning and Motivation*, *89*, 102092. https://doi.org/10.1016/j.lmot.2024.102092