

Establishing CEFR-Referenced Cut Scores for the Royal Police Cadet Academy Test of English Proficiency (RPCA-TEP) Using the Modified Angoff Method

Komjit Chaiyo¹ and Pitupoom Nakprada^{2*}

(Received: November 16, 2025; Revised: January 31, 2026; Accepted: February 11, 2026)

Abstract

The Royal Police Cadet Academy Test of English Proficiency (RPCA-TEP) was designed to evaluate the English language proficiency of police cadets in accordance with the mandate of the Higher Education Commission, which requires English proficiency testing aligned with the Common European Framework of Reference (CEFR). This study aimed to align RPCA-TEP scores with CEFR proficiency levels. Using the Modified Angoff method, 13 subject-matter experts provided judgments on the likelihood of borderline candidates at CEFR levels A2, B1, and B2 answering test items correctly. The analysis established cut-off scores out of 120, with candidates scoring 26–55 classified as A2, 56–82 as B1, and 83–120 as B2. The findings highlight the potential of the RPCA-TEP as a CEFR-referenced assessment tool and recommend further refinement using CEFR ability descriptors to guide future test development.

Keywords: English proficiency test, CEFR-referenced, cut-off scores, Modified Angoff method

Introduction

English proficiency is a core professional competence for modern police officers in Thailand. Officers routinely interact with foreign tourists, expatriates, international organizations, and transnational law-enforcement partners, making the ability to communicate accurately and appropriately in English operationally critical. Miscommunication in police work can have serious consequences, including failure to obtain accurate witness statements, mishandling of victims, and breakdowns in international cooperation (Suwannaphim, 2011; Sae-Jeng, 2018).

¹ Assistant Professor, Department of English, Faculty of Social Science, Royal Police Cadet Academy

² Lecturer, Department of English, Faculty of Social Science, Royal Police Cadet Academy

Research grant from Royal Police Cadet Academy

* Corresponding Author E-mail: pitupoom@gmail.com

To address these challenges, Thailand's higher-education policy requires universities and professional academies to define English proficiency standards using international reference frameworks, particularly the Common European Framework of Reference for Languages (CEFR) (Higher Education Standards Commission, 2024). The CEFR provides a six-level scale (A1–C2) that describes what learners “can do” in real-world communication, allowing institutions to make language requirements transparent, interpretable, and comparable across programs (Council of Europe, 2001).

However, adopting CEFR levels in policy does not automatically ensure that locally developed tests actually measure those levels. CEFR alignment requires that cut scores—the score points used to classify examinees into levels such as A2, B1, or B2—be empirically and theoretically justified. A cut score is defined as a threshold on a test score scale that separates performance categories corresponding to qualitatively different levels of ability (Cizek & Bunch, 2007). Without defensible cut scores, CEFR labels attached to test results become arbitrary and may misrepresent candidates' true communicative competence.

The Need for CEFR-referenced Cut Scores for RPCA-TEP

The Royal Police Cadet Academy (RPCA) uses the Royal Police Cadet Academy - Test of English Proficiency (RPCA-TEP) as one of the instruments for certifying cadets' English readiness for graduation and professional deployment. While cadets may also submit scores from external tests (e.g., TOEIC, IELTS), RPCA-TEP remains the only assessment specifically designed to reflect the linguistic demands of Thai police work. Despite its institutional importance, RPCA-TEP has not yet been formally aligned to CEFR levels through a standard-setting process.

This creates a critical knowledge and policy gap: RPCA currently uses RPCA-TEP scores for high-stakes decisions (graduation and certification), yet there is no defensible basis for interpreting those scores as representing CEFR levels such as A2, B1, or B2. Without CEFR-linked cut scores, it is unclear whether cadets who “pass” RPCA-TEP actually meet the communicative standards required by national higher-education policy or by international policing contexts. This study addresses that gap by establishing CEFR-referenced cut scores for RPCA-TEP.

Alignment of RPCA-TEP Construct with CEFR Descriptors in Police Communication Contexts

To clarify how the RPCA-TEP construct aligns with CEFR proficiency descriptors, Table 1 maps the test sections to CEFR levels A2, B1, and B2 and illustrates how these abilities correspond to communicative tasks encountered in police work. This mapping demonstrates that the RPCA-TEP construct covers receptive language abilities that are directly relevant to operational policing contexts, thereby supporting the rationale for linking RPCA-TEP performance to CEFR levels.

Table 1 Alignment of RPCA-TEP Sections with CEFR Descriptors and Police Communication Tasks

CEFR Level	CEFR Descriptor	RPCA-TEP Sections	Examples of Police Communication Tasks
A2	Can understand and use everyday expressions and simple sentences related to personal information, places, work, and daily life. Can exchange basic information and describe simple background details, needs, and surroundings.	Listening, Vocabulary, Reading	Understanding simple tourist inquiries, reading short notices, identifying personal information from documents, following simple spoken instructions from foreigners
B1	Can understand the main points of clear standard input on familiar matters such as work, school, and leisure. Can handle situations likely to arise while traveling. Can describe experiences, events, and give brief explanations.	Listening, Reading, Language Structure, Vocabulary	Understanding incident reports, reading statements, handling routine interactions with tourists, understanding complaints, explaining procedures in simple English
B2	Can understand extended speech and more complex texts. Can read and comprehend detailed written information. Can communicate fluently and accurately in most situations.	Listening, Reading (longer texts), Language Structure	Understanding detailed reports, reading legal or procedural documents, following complex spoken briefings, interpreting detailed written information from foreign sources

Standard Setting and the Choice of the Modified Angoff Method

This study focuses explicitly on standard setting, not score equating or test linking. Standard setting refers to the process of establishing cut scores that define performance levels on a test (Cizek & Bunch, 2007). Among available standard-setting methods, the Modified Angoff method was selected because it is theory-driven, judgment-based, and widely used for criterion-referenced and certification testing.

In the Modified Angoff procedure, trained subject-matter experts estimate the probability that a minimally qualified candidate at each CEFR level (e.g., A2, B1, B2) would answer each test item correctly. These item-level judgments are aggregated to produce cut scores that are directly anchored in performance expectations rather than statistical distributions (Hambleton & Pitoniak, 2006).

This approach is particularly appropriate for CEFR alignment because CEFR itself is a performance-descriptor framework, not a statistical scale. By asking experts to conceptualize minimally competent CEFR-level candidates and evaluate how they would perform on RPCA-TEP items, the Modified Angoff method operationalizes CEFR descriptors in a defensible psychometric way. Prior research has shown that Angoff-based methods provide reliable and interpretable cut scores when experts are well trained and the construct is clearly defined (Cizek & Bunch, 2007).

Objective

The objective of this study is to establish CEFR-referenced cut scores for RPCA-TEP at the A2, B1, and B2 levels using the Modified Angoff standard-setting method, thereby providing a defensible framework for interpreting RPCA-TEP scores.

Method

1. Population and sample

Thirteen panelists were recruited using purposive, criterion-based sampling, consistent with recommendations that Angoff panels include approximately 10–15 experts to ensure stability of cut-score estimates (Cizek, 2012; Cizek & Bunch, 2007; Tannenbaum & Cho, 2014).

To qualify, panelists were required to: (a) have at least five years of experience teaching English at the tertiary level; (b) be involved in English testing, curriculum design, or test development; (c) demonstrate familiarity with CEFR descriptors; and (d) have experience

with receptive-skills instruction. Preference was given to experts with experience in ESP or policing-related English. Collectively, the panel represented a balance of instructional, assessment, and domain expertise, which is considered essential for defensible standard-setting outcomes.

2. Research instruments

2.1 RPCA-TEP

The RPCA-TEP is a 120-item multiple-choice test comprising Listening (40), Grammar (20), Vocabulary (20), and Reading (40). Approximately 40% of items are in policing-related contexts, with the remainder drawn from general domains.

The test was piloted with 280 fourth-year RPCA cadets, and classical item statistics (item difficulty and discrimination) were computed. These statistics were used to support expert judgment in Round 2 of the Angoff procedure. In line with standard-setting best practice, the pilot data were used to inform but not to override expert judgment (Hambleton & Pitoniak, 2006; Cizek & Bunch, 2007).

Content validity was supported through alignment of test specifications and items with CEFR descriptors and RPCA curriculum objectives, and through expert review during test development.

2.2 Angoff rating materials

A structured rating booklet was developed containing test items, probability-rating fields, CEFR borderline profiles, and rating instructions. The materials were trialed using 10 practice items during training to ensure consistent interpretation of the task.

3. Data collection

Prior to the standard-setting procedure, panelists received training on CEFR descriptors, RPCA-TEP test specifications, and the Modified Angoff method. This was followed by a calibration activity using practice items to ensure a shared understanding of borderline candidate performance at CEFR levels A2, B1, and B2.

The Modified Angoff process was conducted across three rounds. In Round 1, panelists independently estimated the probability that borderline candidates at each CEFR level would answer each test item correctly. In Round 2, panelists were provided with summaries of their initial ratings together with pilot item statistics, including item difficulty and discrimination indices, and were invited to revise their estimates accordingly. In Round 3,

panelists reviewed the aggregated results, engaged in structured discussion to resolve discrepancies, and finalized their judgments.

The inclusion of empirical item statistics in Round 2 follows best practice in data-informed Modified Angoff procedures and has been shown to improve the consistency and realism of expert judgments (Hambleton & Pitoniak, 2006; Cizek & Bunch, 2007; Tannenbaum & Cho, 2014).

4. Data analysis

Final CEFR cut scores were derived from the panelists' Round 3 judgments through a structured aggregation process. For each test item and each target CEFR level (A2, B1, and B2), the probability estimates provided by the 13 panelists were compiled and averaged to obtain an item-level mean probability. This value represented the expected likelihood that a borderline candidate at the specified CEFR level would answer that item correctly.

Because the Angoff judgments addressed the question of what percentage of items borderline candidates would answer correctly, the Round 3 mean cut scores were first expressed as percentages (out of 100). These percentage values were then converted proportionally to the 120-point RPCA-TEP reporting scale to establish the operational cut scores. The CEFR score bands were subsequently defined using these cut scores as boundaries.

Finally, the stability of the standard-setting process was evaluated by examining changes in the dispersion of panelists' estimates across the three rating rounds. Reductions in variability from Round 1 to Round 3 were interpreted as evidence of increasing convergence of expert judgment.

Results

This section presents the outcomes of the Modified Angoff standard-setting procedure conducted across three rounds of expert judgment. The results first demonstrate the convergence of expert estimates across rounds and then present the operational CEFR-linked cut scores derived from Round 3.

Convergence of expert judgments across rounds

Across CEFR levels A2, B1, and B2, a consistent pattern of decreasing standard deviations was observed from Round 1 to Round 3, indicating increasing agreement among panelists as they became familiar with CEFR descriptors, reviewed item statistics, and engaged in structured discussion. Tables 2–4 present the mean estimates and standard deviations for each test section across rounds.

Table 2 Convergence of Expert Judgments Across Rounds for CEFR Level A2

Section	Round 1		Round 2		Round 3	
	Mean	SD	Mean	SD	Mean	SD
Listening	20.39	16.71	18.16	14.03	17.84	12.97
Language Structure	29.21	18.33	26.25	16.18	25.03	15.15
Vocabulary	28.08	15.29	24.78	14.20	24.32	12.10
Reading	22.73	19.36	21.92	18.74	20.36	16.95
Total	25.10	17.42	22.77	15.78	21.88	14.29

Table 3 Convergence of Expert Judgments Across Rounds for CEFR Level B1

Section	Round 1		Round 2		Round 3	
	Mean	SD	Mean	SD	Mean	SD
Listening	47.92	15.42	45.23	14.11	43.59	12.06
Language Structure	54.17	16.33	52.42	14.26	50.26	13.12
Vocabulary	53.10	14.87	51.57	13.74	50.93	11.37
Reading	45.79	17.11	43.63	15.98	43.14	14.53
Total	50.24	15.93	48.21	14.52	46.98	12.77

Table 4 Convergence of Expert Judgments Across Rounds for CEFR Level B2

Section	Round 1		Round 2		Round 3	
	Mean	SD	Mean	SD	Mean	SD
Listening	70.19	16.84	69.24	15.54	67.17	14.15
Language Structure	76.00	15.16	75.53	14.82	73.41	14.33
Vocabulary	73.64	16.58	72.32	14.14	70.68	12.92
Reading	67.38	17.45	66.94	16.01	64.43	15.81
Total	71.80	16.50	71.00	15.12	68.92	14.30

The decreasing standard deviations confirm convergence of expert judgments during the Modified Angoff process. Following this convergence, the Round 3 results were adopted as the operational cut scores.

Operational cut scores from Round 3

Table 5 presents the Round 3 mean percentage cut scores representing the expected proportion of test items that borderline candidates at each CEFR level would answer correctly.

Table 5 Post-discussion Operational Cut Scores from Round 3

CEFR Level	Round 3 Mean Cut Score (%)
A2	21.88
B1	46.98
B2	68.92

Because these values represent percentages (out of 100), they were proportionally converted to the 120-point RPCA-TEP reporting scale to establish operational cut scores, as shown in Table 6.

Table 6 Cut Scores After Proportional Conversion

Cut Score	CEFR Level
26	A2
56	B1
83	B2

These cut scores were then used to define CEFR score bands on the RPCA-TEP scale.

CEFR score bands for RPCA-TEP

Table 7 RPCA-TEP Score Ranges Corresponding to CEFR Levels

RPCA-TEP Score Range	CEFR Level
26 - 55	A2
56 - 82	B1
83 - 120	B2

RPCA-TEP scores within these ranges correspond to CEFR levels A2, B1, and B2 respectively.

Discussion

Interpretation of the CEFR-linked cut scores

The primary objective of this study was to establish CEFR-linked cut scores for the RPCA-TEP using the Modified Angoff method. This objective was successfully achieved. The operational score bands derived from Round 3 were: A2 (26–55), B1 (56–82), and B2 (83–120) on the 120-point RPCA-TEP scale. These score ranges represent the expected performance of

borderline candidates at each CEFR level on the RPCA-TEP. The cut scores therefore provide a meaningful basis for interpreting RPCA-TEP results in relation to CEFR proficiency descriptors, particularly in listening and reading, which constitute the core construct of the test.

Quality and defensibility of the cut scores

The quality of these cut scores is supported by the convergence of expert judgments across the three Angoff rounds. Standard deviations decreased consistently from Round 1 to Round 3 at all CEFR levels, indicating increasing agreement among panelists regarding borderline performance. This convergence reflects effective expert familiarization with CEFR descriptors, the use of empirical item statistics, and structured panel discussion.

Ricker (2006) highlights that iterative rounds, structured panel discussion, and presentation of empirical item statistics are key features that reduce variability among experts and improve agreement in cut-score estimation. The pattern observed in this study therefore aligns closely with these recommended practices in standard-setting research. Such convergence is widely recognized as evidence of stable and defensible cut scores rather than arbitrary estimates.

Uncertainty, fairness, and careful interpretation

Despite their defensibility, these cut scores should be interpreted as decision thresholds rather than precise indicators of overall language ability. Because RPCA-TEP currently assesses only receptive skills, it does not represent the full range of CEFR descriptors, particularly those related to speaking and writing. Users of the test must therefore avoid overgeneralizing that candidates classified at a CEFR level can perform all functions described at that level. Classification uncertainty is greatest for candidates whose scores fall near the cut points, and careful interpretation is required to ensure fairness in decision-making.

Practical use of the score bands at RPCA

The CEFR-linked score bands provide RPCA with a clear framework for interpreting test results. They can support graduation decisions, diagnostic placement, curriculum alignment, and internal policy development within the academy. However, these scores are intended for institutional use and should not yet be treated as directly equivalent to international tests such as IELTS or TOEFL without further external validation.

Limitations

Several limitations should be acknowledged when interpreting the findings of this study. The RPCA-TEP currently assesses only receptive language skills (listening and reading). As a result, the established cut scores reflect CEFR alignment within this limited construct and do not encompass productive skills such as speaking and writing, which are integral to full CEFR proficiency.

The standard-setting panel consisted of 13 experts with relevant experience in English language teaching, assessment, and police education. This panel size is fully consistent with recommended practice in Modified Angoff procedures, where panel quality and contextual expertise are considered more critical than sheer numbers. Importantly, the experts represented diverse professional backgrounds. Future studies may broaden stakeholder representation further, for example by including CEFR specialists or international assessors to enrich perspectives on descriptor interpretation.

The study did not include external validation against international CEFR-benchmarked English proficiency tests (e.g., IELTS, TOEFL) or independent performance indicators. Consequently, while the cut scores are defensible within the RPCA context, additional post-alignment validation is necessary to examine classification accuracy and broader test-use validity.

As with all judgment-based standard-setting procedures, classification uncertainty is greatest for candidates whose scores fall near the cut points. Although the convergence of expert judgments supports the reliability of the established thresholds, careful interpretation remains necessary for borderline cases.

Addressing these limitations through future research will further strengthen confidence in the RPCA-TEP as a CEFR-referenced assessment and support its responsible use in police education

Recommendation

To strengthen the RPCA-TEP and guide future research, several actions are recommended. First, the present study was conducted within a single institution and among cadets with relatively similar academic backgrounds; therefore, future comparative and longitudinal studies across different faculties, cohorts, and police-education contexts are recommended to provide a broader understanding of how English proficiency and learner perceptions develop over time.

Second, for assessment improvement, post-alignment validation procedures should be systematically implemented. These include conducting content audits to ensure alignment with CEFR descriptors specific to policing tasks, examining correlations with external criteria such as international English test scores or supervisor performance ratings, and undertaking studies of decision accuracy and rater reliability to verify the technical soundness of the assessment.

Finally, ongoing attention to fairness, impact, and test-use validity is crucial. It is recommended that the RPCA routinely monitor these components to maintain transparency, support defensible decision-making, and uphold the long-term credibility of the RPCA-TEP.

References

- Cizek, G. J. (2012). *Setting performance standards: Foundations, methods, and innovations* (2nd ed.). Routledge.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Sage.
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.
- Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Praeger.
- Ricker, K. L. (2006). *Setting cut-scores: A critical review of the Angoff and modified Angoff methods*. *The Alberta Journal of Educational Research*, 52(1), 53–64.
- Sae-Jeng, S. (2018). *Problems and needs of the occupational English language skills for Thai tourist police officers working at one of the international airports in Thailand* [Independent study]. Thammasat University.
https://ethesisarchive.library.tu.ac.th/thesis/2018/TU_2018_6021040271_10740_10283.pdf
- Suwannaphim, S. (2011). *English oral communication problems of police officers in high tourism areas* [Master's thesis]. Khon Kaen University.
- Tannenbaum, R. J., & Cho, Y. (2014). Critical thinking and the Angoff method: A study of standard-setting panelists' judgments. *Educational Measurement: Issues and Practice*, 33(2), 14–24. <https://doi.org/10.1111/emip.12033>