การตรวจสอบความเที่ยงตรงของแบบทดสอบภาษาอังกฤษมาตรฐานของวิทยาลัยนอร์ทเทิร์น (NTC-SET) ตามกรอบมาตรฐานความสามารถทางภาษาของยุโรป (CEFR)

ธัญพิสิษฐ์ เป็นใจช่วย [1]

[1] ผู้นิพนธ์ประสานงาน โทรศัพท์ 08-9460-9021 อีเมล : thanphisitjay@gmail.com

รับเมื่อ 23 พฤษภาคม 2568 ตอบรับเมื่อ 24 ตุลาคม 2568 10.14416/j.faa.2026.25.001

บทคัดย่อ

วัตถุประสงค์ของการศึกษานี้คือเพื่อพัฒนาและทวนสอบความตรงของแบบทดสอบวัดความสามารถทางภาษาอังกฤษมาตรฐานของวิทยาลัยนอร์ทเทิร์น (NTC-SET) ที่อ้างอิงตามกรอบมาตรฐาน CEFR การวิเคราะห์นี้มุ่งเน้นไปที่กระบวนการออกแบบข้อสอบ ความตรงเชิงเนื้อหา ความยากง่ายของข้อสอบ ค่าอำนาจจำแนก และความน่าเชื่อถือแบบความสอดคล้องภายใน โดยใช้ข้อมูลจากผู้เข้าร่วมการทดสอบ จำนวน 351 คน ในส่วนของความตรงเชิงเนื้อหา ได้รับการตรวจสอบโดยผู้เชี่ยวชาญ 3 ท่าน ซึ่งให้ค่าดัชนีความสอดคล้องระหว่างข้อสอบกับวัตถุประสงค์ (IOC) ในระดับสูง โดยมีค่าตั้งแต่ 0.76 (การพูด) ถึง 1.00 (การฟัง) การวิเคราะห์ข้อสอบรายข้อพบว่าข้อสอบมีระดับความยากโดยรวมปานกลาง (P = 0.50) และมีค่าอำนาจจำแนกที่ดีมาก (0.43) ซึ่งแสดงให้เห็นว่าข้อสอบสามารถจำแนกระดับความสามารถของผู้สอบได้อย่างมีประสิทธิภาพ สำหรับความน่าเชื่อถือแบบความสอดคล้องภายในซึ่งวัดโดยใช้สูตรคูเดอร์-ริชาร์ดสัน 20 (KR-20) นั้น อยู่ในเกณฑ์ดี โดยส่วนของการอ่านมีค่าสัมประสิทธิ์ความน่าเชื่อถือสูงสุดที่ 0.94 โดยสรุป ผลการวิจัยนี้ถือเป็นหลักฐานเชิงประจักษ์ที่หนักแน่นซึ่งสนับสนุนความตรงและความน่าเชื่อถือของแบบทดสอบ NTC-SET ในฐานะเครื่องมือสำหรับประเมินความสามารถทางภาษาอังกฤษตามกรอบมาตรฐาน CEFR

คำสำคัญ : ความตรงของแบบทดสอบ แบบทดสอบภาษาอังกฤษมาตรฐาน CEFR ความสอดคล้องระหว่างข้อสอบกับวัตถุประสงค์ ความน่าเชื่อถือของแบบทดสอบ

[1] อาจารย์ประจำศูนย์ภาษา มหาวิทยาลัยเนชั่น

Examining the Validity of Northern College Standardized English Test (NTC-SET) according to the Common European Framework of Reference for Languages (CEFR) Standard

Thanphisit Pernjaichuay [1]

[1] Corresponding author Mobile: 08-9460-9021, E-mail : thanphisitjay@gmail.com

Received 23 May 2025; Accepted 24 October 2025 10.14416/j.faa.2026.25.001

Abstract

The purposes of this study were to develop and validate the Northern College Standardized English Test (NTC-SET), a standardized English proficiency test aligned with CEFR benchmarks. The analysis focused on NTC-SET's design process, content validity, item difficulty, item discrimination, and internal consistency reliability. There were 351 participants used in the study. Content validity was confirmed by three expert raters with high Index of Item-Objective Congruence (IOC) ratings ranging from 0.76 (speaking) to 1.00 (listening). The item analysis revealed a moderate overall difficulty level (P=0.50) and a very good discrimination index (0.43), showing that the test successfully distinguished between skill levels. The internal consistency reliability, as measured using the Kuder-Richardson Formula 20 (KR-20), was strong, with the reading section having the highest reliability coefficient, at 0.94. Overall, the findings offered solid empirical evidence supporting the NTC-SET's validity and reliability as a tool for assessing English competence in accordance with CEFR norms.

Keywords : Test Validity, Standardized English Test, CEFR, Item-objective Congruence, Test Reliability

[1] Lecturer, Language Center, Nation University

1. Introduction

The English language plays an essential role as the most common medium for exchanging information across borders, making it a critical component for gaining knowledge, building partnerships, and expanding into new markets. The EF English Proficiency Index (EF EPI) found strong and consistent correlations between English proficiency and various measures of trade, innovation, human capital development, and competitiveness. It can be implied that a person who can use English at a proficient level will have an advantage and be a strong candidate for employment in any company. According to the 2022 EF EPI report, Asia's regional average score decreased slightly, reflecting lower overall proficiency scores from the countries within the region. The average ratings for South Asia and ASEAN remained similar to those of the previous year. At present, only a small proportion of people in Thailand can communicate effectively in English. Despite ongoing efforts, significant gaps in English proficiency remain. Survey results showed that Thailand ranked 97th out of 111 countries, placing it in the "very low proficiency" category (EF Education First, 2022).

Thailand's universities frequently employ well-known standardized tests, such as IELTS, TOEFL, or TOEIC to assess language proficiency. Furthermore, many institutions and colleges create their own institutional competency exams to evaluate students for specific goals like placement or graduation. Recognizing the importance of these, Northern College created a series of exams to assess its students' language abilities in accordance with the Common European Framework of Reference for Languages (CEFR).

However, while many such institutional tests exist, there is a significant research gap in the absence of published studies verifying their validity. Previous research on test validation has been concentrated on large-scale global exams, with less emphasis being placed on regionally created tests in contexts such as Thailand. Without empirical investigation of these regional assessments, they are typically assumed to be appropriate for their intended contexts yet lack objective validation. This research study attempts to address that gap.

Therefore, this research was conducted to formally examine the quality of the Northern College test package by applying principles of measurement and evaluation. This study investigates the test's validity and reliability to determine if it is a psychometrically sound assessment tool. The conclusions of this study could assist Northern College and other educational institutions in developing and evaluating their own English language exams.

2. Research Objective and Questions

The study seeks to evaluate the Northern College Standardized English Test (NTC-SET) by investigating its creation process and psychometric features in connection to the Common European Framework of Reference (CEFR). To achieve this objective, the following research questions were developed :

2.1 What procedures were involved in the development process of the NTC-SET?

2.2 To what extent does the NTC-SET demonstrate content validity in alignment with the CEFR standards?

2.3 What are the item difficulty and item discrimination indices of the NTC-SET?

2.4 What is the internal consistency reliability of the NTC-SET?

3. Literature Review and Conceptual Framework

This section explores the theoretical concepts and empirical findings that serve as the basis for this research. It begins with a brief review of the Common European Framework of Reference for Languages (CEFR), the international standard that provides the foundational blueprint for the NTC-SET. The section continues with an overview of the fundamental principles of test quality, validity, and reliability, as well as the statistical techniques used to evaluate them. It then briefly reviews previous research on test validity in order to lay the groundwork for the current study. Finally, it establishes the conceptual foundation for investigating the Northern College Standardized English Test (NTC-SET).

3.1 The Concept of Test Quality : Validity and Reliability

The quality of any educational assessment is primarily determined by two key principles : validity and reliability. Validity is the extent to which a test measures what it promises to measure, whereas reliability is the consistency with which those measurements are made.

The validity is the most important consideration in test evaluation. It is the degree to which evidence and theory support the interpretations of test scores (AERA, APA, & NCME, 2014). This study investigates content validity, which can be defined as a level at which test items accurately reflect the language skills being evaluated.

The reliability means that a metric is always the same. If a test gives the same results when the conditions are the same, it is said to be reliable. This study examines internal consistency reliability, which evaluates the uniformity of items within the test. This is often evaluated by administering the test once and subsequently using statistical techniques to calculate a reliability coefficient.

3.2 Item Analysis : Difficulty and Discrimination

The quality of a test is determined by how valid and reliable it is throughout, as well as the effectiveness of its individual items. Item analysis is performed to determine the effectiveness of each item. This study focuses on two primary indicators:

Difficulty Index (p): This index is used to evaluate individual test items by measuring the proportion of test-takers who answer the item correctly. A score of 0.90 signifies that 90% of the students got their answer accurately, which makes the question appear easy. Items with p-values between .21 and .80 are usually acceptable.

Discrimination Index (r): This shows how well an item can identify the difference between test-takers who do well and those who do not. The value can fall anywhere between -1.0 and +1.0. A positive value is what is desirable because it means that students who did well on the test were more likely to get the answer right. A discrimination index of .20 or higher is typically considered acceptable.

3.3 Previous Studies on Test Validation

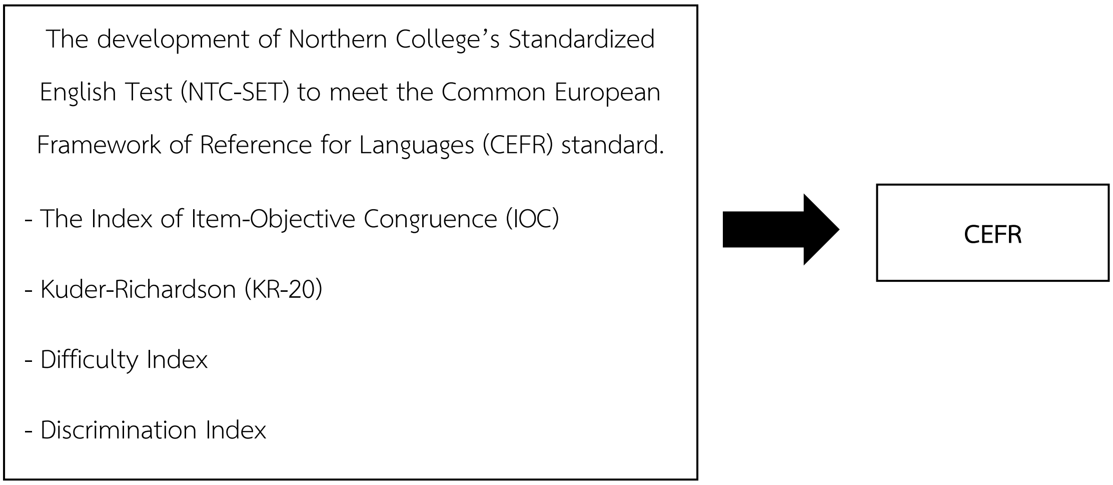
The validation of institutional language tests is a critical area of research, ensuring that local assessments are fair and accurate. For example, Aryadoust and Riazi (2011) evaluated the validity of an Iranian university English placement examination using text analysis and statistical techniques comparable to those used in this study. Their results showed that the test was beneficial to its objective of placing students in groups according to their skill level.

Ratasawang (2020) measured an English proficiency test administered at Thailand's Kasetsart University (KU-EPT) for its validity and item quality. The study found that, while the exam was usually reliable, multiple aspects needed to be strengthened, including low discrimination indices and concerns regarding item complexity.

These studies show the importance of conducting local validation research to ensure the quality of institutional assessments. Nevertheless, there exists a lack of published research concerning the validation of English proficiency tests in private colleges specifically in Northern Thailand, an oversight that this study aims to address.

3.4 Conceptual Framework

Figure 1 depicts the process of validating the NTC-SET and serves as the study's conceptual framework. It is based on the Classical Test Theory (CTT) approach (Allen & Yen, 2002). The framework positions the NTC-SET as the primary subject of investigation. The test is evaluated against the external benchmark of the CEFR standard. This evaluation is conducted using a set of psychometric tools. Specifically, Content validity is assessed using the IOC. Internal consistency reliability is measured by KR-20. The Difficulty Index and Discrimination Index examine the quality of each item. Collectively, the results of these analyses provide the empirical evidence necessary to determine the overall quality and validity of the NTC-SET.



The development of Northern College's Standardized English Test (NTC-SET) to meet the Common European Framework of Reference for Languages (CEFR) standard.

- The Index of Item-Objective Congruence (IOC)
- Kuder-Richardson (KR-20)
- Difficulty Index
- Discrimination Index

CEFR

Figure 1 : Conceptual Framework for the Validation of the NTC-SET

4. Research Methodology

This section describes the research methodology used in this study. It is divided into five sections : the research method, the participants and sampling procedure, the research instruments, the data collection method, and the statistical treatment of data.

4.1 Research Method

This study used a quantitative approach to measure the validity and reliability of Northern College's Standardized English Test (NTC-SET) against the Common European Framework of Reference for Languages (CEFR) standard. The analysis involved several statistical procedures. First, the Index of Item-Objective Congruence (IOC) was employed to determine the test items' content validity. Second, the internal consistency reliability of the exam was assessed using Kuder-Richardson (KR-20). Finally, the Difficulty Index and Discrimination Index were used to assess the quality of each test item.

4.2 Participants and Sampling

4.2.1 Target Participants and Sample Size

The target population for this research comprised students nearing graduation at Northern College. The researcher gathered data from students in their final year, both bachelor's and master's, who had been required to take the NTC-SET as an exit examination. The respondents are from different faculties and represent the usual test-taker population. The date of data collection was February 2024.

The college has a total student population of over 2,000, with approximately 750 graduates each year. The study included 351 participants drawn from this demographic. This sample size was considered sufficient for conducting a valid statistical study of the test's performance. Given a population of approximately 750 graduates, this sample size exceeds the 261 participants required for a 95% confidence level with a 5% margin of error (Yamane, 1967).

4.2.2 Sampling Procedure

A purposive sampling strategy was used in this investigation. The researcher chose individuals from students in their final year of both bachelor's and master's degrees who had formally enrolled to take the NTC-SET during the data collection period. This strategy was adopted to verify that all participants were from the intended test-taking group for which the exam was developed.

4.3 Research Instruments

The NTC-SET was the primary instrument utilized during this study. The instrument was developed and validated in two stages: a pretesting stage for content validation, and the execution of the final test for gathering data.

4.3.1 The NTC-SET Examination

The NTC-SET is a 100-item multiple-choice exam designed to assess English ability across the whole Common European Framework of Reference for Languages (CEFR), from A1 (Basic User) to C2 (Proficient User) (Council of Europe, 2001). To achieve both functional competence and linguistic control, all items were constructed using fundamental CEFR alignment concepts. Each item was prepared to test a specific communicative activity defined in the CEFR's "Can-Do" assertions, with items tailored to the various abilities at each level (A1-C2). Furthermore, the language used in test items and passages was carefully selected to correspond to the vocabulary and grammatical structures necessary for each CEFR level, using resources such as the English Vocabulary Profile (EVP) (Cambridge University Press, n.d.). This assures that lower-level items use basic language, while higher-level items use more complicated and nuanced vocabulary. The alignment of these design principles with the CEFR was subsequently verified using an expert validation procedure, as described in the following section. The test is divided into four sections:

- Semi-Speaking (25 items) : This section evaluates students' ability to respond appropriately in a variety of conversational scenarios, including ordinary life, work, and formal settings. These scenarios were developed to demonstrate the B1 Can-Do statement : 'Can enter unprepared into conversation on subjects that are familiar, of personal interest, or pertinent to everyday life (e.g., family, hobbies, job, travel, and current events)' (Council of Europe, 2001, p. 26). Multiple-choice questions allow test takers to choose the most pragmatically appropriate response.
- For example : **Question :** How did the meeting with the safety inspectors go?
  (A) They all came by car.
  (B) The safety inspectors go to the site.
  (C) We have to be ready at the site by ten.
  (D) It was very interesting to hear their opinions.
- Listening (25 items) : This section assesses comprehension of spoken English from a variety of sources, including news, announcements, and lectures. The tasks are aligned with the B2 Can-Do statement : 'Can understand the main ideas of complex speech on both concrete and abstract topics, including technical discussions in their field of specialization' (Council of Europe, 2001, p. 27).
- For example : **Question :** How can someone combat the negative thoughts associated with stage fright?
  (A) By breathing shallowly.
  (B) By staring at the audience.
  (C) By focusing on what needs to be done.
  (D) By thinking about possible negative outcomes.
- Semi-Writing (25 items) : This section multiple-choice questions tests whether students comprehend how to apply proper language in writing, particularly in formal emails and messages at work. This aligns with the B1 descriptor for written production : 'Can write straightforward connected texts on topics which are familiar or of personal interest, focusing on grammatical accuracy and appropriate register in common workplace contexts' (Council of Europe, 2001, p. 26).

- For example : **Question :** _________, domesticated grapes grow in clusters, range in color from pale green to black, and contain sugar in varying quantities.
  (A) Their botanical classification as berries
  (B) Although their botanical classification as berries
  (C) Because berries being their botanical classification
  (D) Because they are botanically classified as berries
- Reading (25 items) : This section evaluates understanding of diverse text types, such as short passages, formal emails, and reports. The passages and questions were developed to assess the B2 Can-Do statement: 'Can read articles and reports concerned with contemporary problems in which the writers adopt particular attitudes or viewpoints.' (Council of Europe, 2001, p. 27).
- For example : **Question:** What is the article mainly about?
  (A) New jobs.
  (B) Real estate values
  (C) A development project
  (D) Environmental protection

4.3.2 Pretesting and Instrument Validation

The pretest was conducted in order to ensure the content validity of all test questions. A pilot version of the 100-item exam was distributed to a panel of three professional judges with experience in English language evaluation and the CEFR.

The experts evaluated each question using the Index of Item-Objective Congruence (IOC), assigning a score of +1 (clearly measures the objective), -1 (does not measure the objective), or 0 (unsure). As indicated in Table 1, the total IOC value for the test was 0.87, with all parts scoring much higher than the permitted threshold of 0.50. This result indicated that the instrument had strong content validity.

Table 1 : Pretesting Item Objective Congruence (N = 3)

Variables	Number of Items	IOC Value
Semi-Speaking	25	0.76
Listening	25	1
Semi-Writing	25	0.83
Reading	25	0.89
Overall	100	0.87

4.4 Data Collection Method

The test was administered to the 351 student participants through online platforms. The university’s official School Information System Advance (SISA) was used for the main administration of the test. The adoption of an online platform was justified for numerous reasons: it allowed for standardized exam conditions for all students, provided a successful method for gathering data from a large number of participants, and accommodated students who were unable to attend classes on campus. During the pretesting phase, Google Forms was utilized to disseminate the exam to the expert judges for ease of use and feedback.

4.5 Data Analysis

The collected data was analyzed using several statistical methods. IOC was used in the pretesting phase to confirm content validity. For the main data which came from the student responses, KR-20 was used to determine the internal consistency reliability of the test. Finally, the Difficulty Index and Discrimination Index were calculated for each item to analyze its individual quality and effectiveness.

To evaluate content validity, this study uses the Index of Item-Objective Congruence (IOC). The IOC was introduced by Rovinelli and Hambleton (1977) as a procedure for evaluating content validity during item development. At least three experts evaluate each question to see if it correlates with the goals of the test. They provide ratings as follows: +1 if the question measures the objective, 0 if unsure, and -1 if it does not. Questions with an IOC value between 0.50 and 1.00 are considered to have acceptable content validity.

This study employs KR-20 to assess internal consistency, Kuder-Richardson 20 (KR-20) is used to measure the internal consistency of a test where questions are scored as either right or wrong (dichotomous). The KR-20 technique considers both the proportion of correct and incorrect answers for each item and the overall score variance. A KR-20 grade of more than 0.60 is generally considered satisfactory, with numbers closer to 1.0 indicating higher reliability.

Difficulty Index (p) is used to evaluate individual test items by measuring the proportion of test-takers who answer the item correctly. A score of 0.90 signifies that 90% of the students got their answer accurately, which makes the question appear easy. Items with p-values between .21 and .80 are usually acceptable. Discrimination Index (r) shows how well an item can identify the difference between test-takers who do well and those who do not. The value can fall anywhere between -1.0 and +1.0. A positive value is what is desirable because it means that students who did well on the test were more likely to get the answer right. A discrimination index of .20 or higher is typically considered acceptable.

5. Results

This section presents the results of the data analysis. It starts with a descriptive study of the participant demographics, then proceeds to the NTC-SET reliability and item analysis results.

5.1 Descriptive Analysis of Participants

The study was conducted with a total of 351 students nearing graduation at Northern College. The demographic characteristics of the participants, including gender, age, and education level, are summarized in Table 2 below.

Table 2 : Demographic Characteristics of Participants (N = 351)

Characteristic	Category	Frequency (N)	Percentage (%)
Gender	Female	205	58.3
	Male	146	41.7
	Total	351	100
Age Group	21 - 30	129	36.67
	31 – 40	76	21.67
	41 - 50	105	30
	51 - 60	41	11.66
	Total	351	100
Education	Below bachelors	327	93.2
	Bachelors	24	6.8
	total	351	100

As the table shows, the participant group consisted of students across a range of ages, with the largest group being between 21-30 years old.

5.2 Reliability Analysis

The researcher applied the Kuder-Richardson Formula 20 (KR-20) to assess the internal consistency reliability of the exam. This method is appropriate because the test items are scored dichotomously, meaning each answer is marked as either correct or incorrect.

The overall reliability coefficient (KR-20) of the 100-item NTC-SET was 0.94. According to Kuder and Richardson's (1937) recommendations, a reliability rating of 0.70 or greater is regarded as appropriate for this test. The achieved result of 0.94 indicates that the NTC-SET has excellent internal consistency. This means that the test items accurately reflect the same underlying English proficiency skills.

Table 3 : Reliability Statistics for the NTC-SET (N = 351)

Variables	Number of Items	KR-20
Semi-Speaking	25	.78
Listening	25	.77
Semi-Writing	25	.76
Reading	25	.83
Overall	100	.94

5.3 Item Analysis

The level of accuracy of the individual items was determined using two essential metrics : the Difficulty Index and Discrimination Index.

5.3.1 The Difficulty Index

The Difficulty Index (P) indicates the proportion of test-takers who answered an item correctly. The P-value for each item can range from 0 (very difficult) to 1 (very easy). The test Difficulty Index was equal to .5051 for all 100 items.

This average value falls within the ideal range of .20 to .80, as suggested by Chaiyasuparakul (1999). This indicates that, overall, the NTC-SET has a moderate level of difficulty, which is appropriate for its target students. An analysis of individual items showed that 95% of the questions were within this acceptable range.

Table 4 : Summary of Item's Difficulty Index

Variables	Number of Items	P-Value
Semi-Speaking	25	.53
Listening	25	.49
Semi-Writing	25	.46
Reading	25	.54
Overall	100	.51

5.3.2 The Discrimination Index

The Discrimination Index (r) measures how well a test item differentiates between high-scoring and low-scoring test-takers. The value ranges from -1.0 to +1.0. According to Ebel (1972), an *r*-value of .20 or higher is considered acceptable.

The average Discrimination Index for the NTC-SET items was .4325. This result is considered very good, suggesting that the test questions effectively distinguish between students with higher and lower English proficiency levels.

Table 5 : Summary of Item's Discrimination Index

Variables	Number of Items	R-Value
Semi-Speaking	25	.38
Listening	25	.47
Semi-Writing	25	.39
Reading	25	.50
Overall	100	.43

6. Conclusion and Discussion

This section highlights the study findings, demonstrating that they have an effect on the study topic and the wider picture of English language evaluation.

6.1 Conclusion

The study was conducted to evaluate the psychometric properties of the NTC-SET using data from 351 graduating students. The main findings are as follows:

- Content Validity : The test has strong content validity, as validated by expert judges, with an overall Index of Item-Objective Congruence (IOC) of 0.87.
- Reliability : The test's internal consistency reliability was high, with a Kuder-Richardson 20 (KR-20) coefficient of 0.94.
- Item Difficulty : The difficulty index is equal to 0.5051, which is considered to be at a moderate level.
- Item Discrimination : The test has a significant level of discriminatory power, with an average discrimination index of 0.4325.

In conclusion, the statistical analysis showed that the NTC-SET is a statistically reliable tool.

6.2 Discussion

The findings of this study provide solid evidence that the NTC-SET is valid and reliable as an institutional English proficiency test linked with CEFR criteria. The high overall IOC value (0.87) is an important finding since it demonstrates that the exam content in the experts' views, is directly relevant to the skills and standards described in the CEFR. This can be fundamental support for the validity of the test.

Moreover, the excellent internal consistency reliability (0.94) suggests that the test items all assess the same underlying concept of English proficiency. The test demonstrates consistent and stable results, with the reading section showing the highest reliability compared to the other components. The reading section was more objective and less challenging than the other skills, which possibly led to more consistent results.

The overall item statistics reflect the test's quality. The moderate Difficulty Index (0.5051) indicates that the test's difficulty is well-matched to the expected proficiency level of the graduating students, reducing the risk of floor or ceiling effects. Crucially, the strong Discrimination Index (0.4325) shows that the test is effective at its main purpose: differentiating between students who have met the required competence level for graduation and those who have not. This is essential for a valid exit examination.

Table 6 : Summary of Items Testing

Analysis Tools	Value
Reliability	.94
Difficulty Index	.51
Discrimination Index	.43

7. Recommendations

The last section concludes the primary findings of the research in accordance with the research objectives. It also considers the significant effects of the findings, limitations of the study, as well as makes recommendations for further research.

7.1 Implications of the Study

The goal of this study was to validate the Northern College Standardized English Test (NTC-SET) as an instrument aligned with the Common European Framework of Reference for Language. Through an analysis of the test's content validity (IOC), internal consistency reliability (KR-20), item difficulty, and item discrimination, the study found strong evidence that the NTC-SET is a psychometrically sound instrument. It is evident from the results that the test is capable of accurately and consistently assessing the English language proficiency of the graduating students at Northern College.

The findings of this research have several important implications:

- For Northern College : The most obvious implication is that Northern College can confidently use the NTC-SET as a valid and reliable exit test. The findings provide a definitive baseline for student performance, which may be used to inform curriculum development and teaching methods.
- For other institutions : This study provides a viable blueprint for other Thai colleges and those across the wider region to follow when conducting validation studies for their own locally produced examinations. It demonstrates that institutional tests, when rigorously evaluated, can be effective and appropriate for their specific contexts.
- For English language assessment : The findings emphasize the importance of institutional-level assessment. Assessment is becoming increasingly important worldwide; thus, this study demonstrates that well-designed local tests can provide a more targeted and accessible way of monitoring student learning outcomes in accordance with global criteria such as the CEFR.

7.2 Limitations of the Study

The researcher acknowledges the following limitations:

- Sample generalizability : The study's participants were all from Northern College. The results showed that the findings on test performance may not be generalizable to students at other institutions with different academic backgrounds or curriculum.
- Test administration mode : The NTC-SET was only administered via an online platform. This study did not look into any disparities in performance that could arise if the test was given in a typical paper-based format.

7.3 Suggestions for Future Research

Based on the findings and limitations of this study, the following areas are recommended for future research:

- The researcher may consider gathering data from different samples, such as multiple institutions across Thailand, in order to achieve greater generalizability.
- A comparative study could be designed to compare students' performance on the online version of the NTC-SET to their performance on the paper-based version.
- Future studies may examine the validity of the NTC-SET by tracking student scores and comparing them with future academic performance as well as job outcomes.

8. Reference

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory*. Waveland Press.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Aryadoust, V., & Riazi, A. M. (2011). A study of the construct validity of a university English placement test. *International Journal of Language Testing*, *1*(1), 58–86.

Cambridge University Press. (n.d.). *English Vocabulary Profile*. Retrieved from http://www.englishprofile.org/wordlists

Chaiyasuparakul, S. (1999). *การทดสอบและวัดผลการศึกษา* [Testing and educational measurement]. Prakaipreuk Publishing House.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.

Ebel, R. L. (1972). *Essentials of educational measurement*. Prentice-Hall.

EF Education First. (2022). *EF English Proficiency Index: A ranking of 111 countries and regions by English skills*. https://www.ef.com/wwen/epi/

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151–160.

Ratasawang, P. (2020). An item analysis of KU-EPT Reading Test. *Journal of Education and Human Development*, *9*(2), 65–72.

Rovinelli, R. J., & Hambleton, R. K. (1977). *On the use of content specialists in the assessment of criterion-referenced test item validity*. Laboratory of Psychometric and Evaluative Research, University of Massachusetts.

Yamane, T. (1967). *Statistics : An introductory analysis*. 2nd ed. Harper and Row.