

CHAPTER 1

INTRODUCTION

A. Background of Study

In today's globalized world, English proficiency has become an essential skill in both academic and professional settings. This is particularly true in non-native English-speaking countries such as Indonesia, where English is categorized as a foreign language. As part of the *Expanding Circle* described by Kachru (1992), Indonesia has integrated English into its national education system from primary school to university level. English mastery is increasingly viewed as an important competency for accessing global knowledge, competing in the job market, and participating in international academic and professional activities.

In response to this growing demand, universities in Indonesia, including Universitas Nahdlatul Ulama Purwokerto (UNU Purwokerto), have implemented English Proficiency Tests (EPTs) as formal tools to assess students' English ability. These tests serve various academic purposes, such as academic learning process and fulfillment of graduation requirements. The use of EPTs reflects the need for standardized and objective measurements of students' English proficiency in higher education.

UNU Purwokerto has developed its own institutional English test known as the English Proficiency Test of Universitas Nahdlatul Ulama (EPTUNU). The test consists of three sections: Listening Comprehension, Structure and Written Expression, and Reading Comprehension. The university sets a minimum passing score of 400 for most study programs, while students in the English Language Education Study Program are required to achieve a higher score of 450. These score requirements are intended to ensure that graduates possess sufficient English skills to support their academic study and future professional careers.

However, as a locally developed test, EPTUNU needs to be evaluated to ensure that it functions effectively as a measurement tool. A good language

test should accurately measure students' abilities, produce consistent results, and fairly distinguish between students of different proficiency levels. According to Bachman and Palmer (2010), test quality is closely related to how well its items are designed and how accurately scores are interpreted. Without proper evaluation, there is uncertainty as to whether EPTUNU truly measures students' English proficiency or merely reflects test-taking strategies or surface-level performance.

In higher education, the quality of proficiency tests is especially important because test results are used for important academic decisions such as placement and graduation (Kunnan & Wu, 2010). In Indonesia, many universities face challenges in developing high-quality English proficiency tests due to limited use of systematic test analysis. As Widodo (2016) notes that English assessments in Indonesian universities often emphasize receptive skills, particularly reading and grammar, but are not always supported by strong psychometric evaluation. Similarly, Istiqlal et al. (2025) report that insufficient test analysis frequently results in items with unbalanced difficulty levels and weak ability to differentiate between high- and low-proficiency students.

To address these issues, modern measurement approaches such as Item Response Theory (IRT) are widely used to evaluate test quality. IRT analyzes the relationship between students' ability levels and their responses to test items, allowing researchers to examine item difficulty and item discrimination more accurately (Embretson & Reise, 2013). One commonly used IRT-based software is the QUEST program, developed by Adam & Khoo (1996). QUEST has been applied in various language testing studies to identify problematic items and improve test validity and reliability (Aryadoust, 2019).

This study focuses on the Reading Comprehension section of EPTUNU, as it is the most suitable component for analysis using QUEST. Reading comprehension items are typically presented in multiple-choice format with clear questions and right or wrong answers, which fits well with the Rasch model applied in QUEST (Wijayanto, 2024). In contrast, Listening Comprehension items may be influenced by external factors such as audio

clarity, test room conditions, and students' concentration levels, which can affect the accuracy of measurement. Meanwhile, Structure and Written Expression items often measure isolated grammatical knowledge and may encourage guessing or test-taking strategies rather than reflecting overall language ability. Reading comprehension, on the other hand, is a core academic skill that reflects students' ability to understand academic texts, making it a stable and appropriate focus for psychometric analysis (Grabe, 2009)

To evaluate the quality of EPTUNU, this study examines four main test characteristics: validity, reliability, item difficulty, and item discrimination. Validity refers to whether the test measures what it is intended to measure, while reliability concerns the consistency of test results (Hughes, 2005). Item difficulty shows how easy or difficult a test item is for students, and item discrimination indicates how well an item distinguishes between students with high and low English proficiency (Bond & Fox, 2015). These four characteristics are essential indicators of test quality and are commonly used in evaluating English proficiency tests (Brown, 2016). Since EPTUNU functions as an institutional test that affects students' academic progress and graduation, it is important that these characteristics are examined carefully.

Several previous studies have analyzed the psychometric quality of institutional English proficiency tests in Indonesian universities. Wijayanto (2024) found that although many items in an institutional EPT showed acceptable difficulty levels, some items had low discrimination and required revision. Similarly, Dewi et al. (2023) reported that variations in item difficulty and discrimination influenced the effectiveness of a university-based English proficiency test. However, these studies were conducted in different institutional contexts and did not specifically examine EPTUNU. Therefore, there remains a research gap in evaluating the psychometric characteristics of EPTUNU using IRT through the QUEST program.

Considering the important role of EPTUNU in academic placement, progression, and graduation at UNU Purwokerto, the absence of a formal psychometric evaluation presents a significant concern. Therefore, this study

applies Item Response Theory using the QUEST program to analyze the Reading Comprehension section of EPTUNU in terms of validity, reliability, item difficulty, and item discrimination. The findings are expected to provide empirical evidence that can support test improvement and strengthen the quality assurance of English proficiency assessment at UNU Purwokerto.

B. Research Questions

1. How good is the validity of the Reading Comprehension section of the English Proficiency Test (EPTUNU) when analyzed using Item Response Theory (IRT) through the QUEST program?
2. How high is the reliability of the Reading Comprehension section of the English Proficiency Test (EPTUNU) based on Item Response Theory (IRT) analysis using the QUEST program?
3. How ideal is the item discrimination of the Reading Comprehension items in the English Proficiency Test (EPTUNU) as measured through Item Response Theory (IRT) using the QUEST program?
4. How ideal is the item difficulty level of the Reading Comprehension items in the English Proficiency Test (EPTUNU) based on Item Response Theory (IRT) analysis using the QUEST program?

C. Objectives of the Study

This study aims to conduct a comprehensive psychometric evaluation of the English Proficiency Test of UNU Purwokerto (EPTUNU) by applying Item Response Theory (IRT) through the QUEST program. Specifically, the objectives of this research are:

1. To determine how fit the validity of the Reading Comprehension section of the English Proficiency Test (EPTUNU) is when analyzed using Item Response Theory (IRT) through the QUEST program.
2. To determine how high the reliability of the Reading Comprehension section of the English Proficiency Test (EPTUNU) is based on Item Response Theory (IRT) analysis using the QUEST program.

3. To examine how ideal the item discrimination of the Reading Comprehension items in the English Proficiency Test (EPTUNU) is as measured through Item Response Theory (IRT) using the QUEST program.
4. To analyze how ideal the item difficulty level of the Reading Comprehension items in the English Proficiency Test (EPTUNU) is based on Item Response Theory (IRT) analysis using the QUEST program.

D. Benefits of the Study

This research holds substantial implications for multiple stakeholders in English language education and assessment. By employing the QUEST program to analyze the English Proficiency Test (EPT) at Universitas Nahdlatul Ulama Purwokerto, the study offers valuable insights that can enhance testing practices and language education policies.

1. For Policymakers

The findings provide policymakers with empirical evidence to guide decisions regarding English language assessment standards in Indonesian higher education. As Kunnan and Wu (2010) notes, the quality of language tests directly impacts educational outcomes and opportunities. This study's psychometric analysis can inform the development of national guidelines for locally-developed proficiency tests, ensuring they meet international measurement standards. The results may also influence resource allocation for test development and teacher training programs, addressing what Bachman and Palmer (2010) identify as the critical need for assessment literacy among educational administrators.

2. For University Administrators

The university stood to benefit significantly from this systematic evaluation of its English Proficiency Test (EPT). As emphasized by Brown (2016), institutional tests that undergo regular validation maintain their credibility and usefulness. The findings of this study enabled administrators to identify and revise problematic test items, improve the test's overall reliability and validity, enhance the institution's reputation for quality

assessment, and make data-driven decisions regarding test revisions or replacements. This aligns with Hambleton et al. (2011), assertion that IRT analysis provides institutions with sophisticated tools for maintaining test quality.

3. For Lecturers and Test Developers

The detailed item analysis offers practical benefits for English language instructors. As Aryadoust (2019) observes, understanding item-level performance helps educators create more effective assessments. Lecturers can use the findings to: refine their test construction techniques, better align test items with learning objectives, develop more accurate placement decisions, provide targeted feedback to students. This addresses what Huang & Flores (2018) describe as the persistent gap between teaching objectives and assessment practices in many EFL contexts.

4. For Students

The improved test quality directly benefits students by ensuring fair and accurate assessment of their English proficiency. As DeMars (2010) notes, well-constructed tests provide students with valid demonstrations of their abilities. Specific advantages include: more precise placement in appropriate courses, fairer evaluation for graduation requirements, increased confidence in test results and better preparation for international standardized tests. This is particularly important in light of Widodo's (2016) finding that assessment quality significantly affects students' motivation and learning outcomes.

5. For Researchers

The study contributes to the growing body of research on language assessment in Indonesian higher education. As Embretson and Reise (2013) suggest, localized applications of IRT advance our understanding of test performance across diverse contexts. The research: provides a model for institutional test validation, expands the literature on QUEST program applications, offers comparative data for similar studies and identifies areas

for further investigation. This addresses Renandya & Widodo (2016) call for more context-specific research on language assessment in Southeast Asia.

E. Scope of Research

This study was delimited to a quantitative psychometric analysis of the English Proficiency Test administered at Universitas Nahdlatul Ulama Purwokerto (EPTUNU). The population of the study consisted of 500 students who had taken the EPTUNU during the 2025 academic year. From this population, a sample of approximately 10% ($n = 55$ students) was selected. The sample was drawn from EPTUNU test administrations conducted during August, September, and October 2025, ensuring representation across multiple testing sessions and reducing potential bias associated with a single administration.

The determination of the sample size followed the guideline proposed by Arikunto (2006), who suggests that when the population size exceeds 100 participants, a sample ranging from 10% to 15% of the population is considered adequate for quantitative research. Furthermore, a sample size of around 50 examinees is acceptable for Rasch-based Item Response Theory (IRT) analysis, as stable item parameter estimation can be achieved with relatively small samples compared to more complex IRT models (Bond & Fox, 2015).

The primary focus of this study was to evaluate the psychometric characteristics of the EPTUNU items, specifically item difficulty, item discrimination (as inferred through fit statistics), validity, and reliability, by applying Item Response Theory (IRT) as operationalized through the QUEST program. The analysis was conducted using test response data from the selected cohort, allowing for a focused and manageable investigation of the measurement properties of the test instrument.

This research concentrated exclusively on the multiple-choice section of the EPTUNU, namely the Reading Comprehension component, as this item format was the most appropriate for IRT-based calibration using QUEST. Multiple-choice items employ dichotomous scoring (correct/incorrect), which

aligns well with the assumptions of the Rasch model implemented in QUEST (Bond & Fox, 2015). In addition, reading comprehension sections in TOEFL-like assessments typically contain a sufficient number of items to support stable parameter estimation and facilitate the examination of item functioning across different proficiency levels (Hambleton, Swaminathan, & Rogers, 2011).

Moreover, reading multiple-choice items minimize rater effects and scoring subjectivity, ensuring that item difficulty and discrimination are attributable to the characteristics of the items themselves rather than to scorer variability. This objectivity makes reading comprehension items particularly suitable for psychometric evaluation and test refinement purposes (Bachman & Palmer, 2010).

For these reasons, the Reading Comprehension component was identified as the most suitable section for IRT analysis using the QUEST program, whereas test sections involving subjective judgment or complex scoring rubrics, such as writing or speaking fell outside the operational scope of the selected analytical model. This study was concerned solely with the statistical attributes of the test instrument and did not examine external variables such as test-taker anxiety, instructional practices, curriculum design, or broader language acquisition processes. Consequently, the findings and recommendations of this study were confined to providing data-driven insights for the refinement of the EPTUNU item bank and the enhancement of its overall construct validity and measurement quality.