CHAPTER II
LITERATURE REVIEW

A. Language Testing

Language testing can be defined as a structured process aimed at evaluating linguistic ability for purposes such as certification, placement, or decision-making. According to Owen (2023), it involves the systematic collection of information from learners in order to measure their level of language proficiency or track their developmental progress. The scope of language testing extends along a continuum, beginning with informal activities in the classroom, such as self-evaluations and peer feedback, and extending to formal procedures involving standardized tests. High-stakes assessments, in particular, are typically managed by authorized testing organizations, and the resulting certificates are widely recognized for academic admissions or employment selection (Owen, 2023).

A key distinction exists between the concepts of assessment and testing, particularly in terms of their breadth and level of formality. Assessment is a broad term that encompasses all forms of evaluative activities conducted to monitor or measure learning outcomes. In contrast, testing refers more specifically to standardized instruments that are designed to generate consistent and comparable data across test-takers. Within educational practice, such standardized tests serve two important roles: they provide evidence for evaluating the effectiveness of instructional programs, and they establish benchmarks for measuring individual learners' achievement (Bachman & Palmer, 2010).

B. Standardized Test

A test is generally regarded as an instrument for measuring an individual's ability, knowledge, or performance (Fauziati et al., 2014). In the context of language learning, a test functions as a means to evaluate learners' linguistic competence and the extent of their mastery of specific skills.

Mcnamara (2000) defines a language test as a procedure designed to elicit evidence of a learner's capacity to use the language in tasks that reflect communicative ability. Similarly, Mardapi (2012) emphasizes the social responsibility of language testing, arguing that it must be aligned with learners' real-world language use. According to Heaton (1990), the primary purpose of administering a language test is to determine the degree to which learners have mastered the language areas and skills that have been taught.

The application of language testing varies depending on its purpose. McNamara (2000) distinguishes between two broad categories of tests: achievement and proficiency. An achievement test is typically administered at the end of a course or academic year to measure the extent to which learners have achieved the objectives of the curriculum. In contrast, a proficiency test is designed to evaluate whether learners possess the level of language competence required to perform successfully in a specific context, such as pursuing higher education or entering the workforce (Heaton, 1990). Unlike achievement tests, proficiency tests do not necessarily relate to a specific syllabus but instead assess broader language ability and potential future performance.

In addition to this classification, Harmer (2007) identifies four main functions of testing:

1) Placement test. It is a test that places the students at an appropriate level in a program or a course. This test is designed and given in order to use the information of the student's knowledge for putting the students in to groups according to their level of the language.
2) Diagnostic test. It is used to expose learner difficulties, gaps in their knowledge and skill deficiency during a course. This test is meant to provide a documentation of the students' general knowledge at the beginning of the study year for the teachers to plan further work and design an appropriate syllabus for their students.
3) Progress or achievement test. This test is designed to measure learners' language and skill progress in relation to the syllabus that they have been

following. Achievement or progress test is often written by teachers and given to students every few weeks to see how well they are doing.

4) Proficiency test. It is a test which measures how much a person knows or has learnt a language. This test is intended to check the learners' language competence. People use this test if they want to be admitted to a foreign university, get a job or obtain some kind of certificate. The most popular tests over the world are TOEFL and IELTS.

1. English Proficiency Test

A test can be defined as a systematic procedure used to measure an individual's knowledge or ability in a specific domain. Brown (2004) describes a test as a method designed to measure a person's competence within a particular area, while Azhari & Sahputri (2022) view a test as a set of questions or exercises functioning as a tool to assess individual achievement and capability. In educational contexts, tests play a crucial role in evaluating learners' understanding and providing evidence of learning outcomes through structured and objective measurement.

In English language education, one of the most widely used instruments is the English Proficiency Test (EPT), which aims to measure an individual's ability to use English for comprehension and communication across various skills. English proficiency refers to the ability to use English at a certain level of accuracy and fluency, encompassing skills such as reading, writing, listening, and speaking (Yusmalinda, 2023). Proficiency is typically assessed through standardized testing procedures to ensure consistency and comparability of results.

Internationally, the most well-known English proficiency test is the Test of English as a Foreign Language (TOEFL), developed by the Educational Testing Service (ETS) and first introduced in 1963. Over time, TOEFL has evolved into several formats, including the Paper-Based Test (PBT), Computer-Based Test (CBT), and Internet-Based Test (iBT). Other widely recognized English proficiency tests include IELTS, PTE Academic,

and Cambridge English assessments, all of which serve academic and professional purposes worldwide (Djamereng et al., 2021).

At Universitas Nahdlatul Ulama Purwokerto, English proficiency is measured through the English Proficiency Test of UNU (EPTUNU), which is a compulsory graduation requirement. EPTUNU adopts a TOEFL-like format administered locally by the university. The test consists of three sections: Listening Comprehension, Structure and Written Expression, and Reading Comprehension. Unlike official TOEFL tests developed by ETS, TOEFL-like tests are institutionally developed and may adapt items from existing TOEFL materials while maintaining alignment with international testing standards (Fatimatazzahro, 2022).

Among various test formats, multiple-choice items are widely used in English proficiency testing and are considered highly suitable for large-scale assessment and psychometric analysis. Multiple-choice tests allow for objective scoring, minimize rater bias, and provide consistent measurement across examinees (Brown, 2004). From a measurement perspective, multiple-choice items with dichotomous scoring are particularly compatible with Item Response Theory (IRT) and the Rasch model, as they enable stable estimation of item difficulty and model fit (Hambleton et al., 2011). For these reasons, multiple-choice tests are regarded as one of the most efficient and reliable formats for evaluating language proficiency and conducting empirical test analysis.

2. TOEFL as Proficiency Test

Standardized tests are widely used worldwide to measure students' mastery of academic competencies, determine graduation requirements, and regulate entry into higher levels of education (Brown & Abeywickrama, 2010). One widely recognized form of language competence assessment is language proficiency testing, with the Test of English as a Foreign Language (TOEFL) being among the most popular instruments globally.

TOEFL is a standardized test that is administered, scored, and interpreted using consistent procedures (NCTE, 2014). It is designed to

measure the English proficiency of non-native speakers in academic contexts, particularly their ability to use English in university-level settings (ETS, 2016). Introduced in 1963, TOEFL has evolved through several formats, including the Paper-Based Test (PBT), Computer-Based Test (CBT), and Internet-Based Test (iBT). Currently, ETS offers the PBT and iBT formats (Hill & Liu, 2012).

In Indonesia, TOEFL is administered not only by official testing institutions but also by universities through TOEFL-like or institutional tests, which are used for internal academic purposes such as admission or graduation requirements (Mahmud, 2014). Many Indonesian universities adopt TOEFL scores as exit requirements for undergraduate and graduate programs. At Universitas Nahdlatul Ulama Purwokerto, for example, students are required to achieve a minimum TOEFL score ranging from 400 to 450, with English Language Education students required to obtain at least 450 prior to their thesis examination. Similar policies are implemented at other Indonesian universities, such as Universitas Airlangga and Universitas Negeri Surabaya (Musahadah, 2015).

Due to institutional facilities and testing conditions, Universitas Nahdlatul Ulama Purwokerto administers a TOEFL-like English Proficiency Test, known as EPTUNU, to assess students' English proficiency. EPTUNU is a locally developed test that adopts the general structure of the TOEFL Paper-Based Test (PBT) and evaluates three skills Listening Comprehension, Structure and Written Expression, and Reading Comprehension using multiple-choice items. Although EPTUNU is not an official TOEFL test regulated by ETS, its TOEFL-like format aligns with widely accepted proficiency testing practices, leading many Indonesian universities to use such assessments as standard indicators of students' ability to understand and use English in academic contexts.

From this information, it can be concluded that the TOEFL test has been recognized all over the world. By establishing the score based on the format of the test. Therefore, colleges and universities in Indonesia also take

a decision to place the TOEFL test as an indicator of people who are able to use and understand English and ETS has been appointed IIEF (Indonesian International Education Foundation) as its representation in Indonesia (LC Kampung Inggris, 2016).

C. English Proficiency Test of UNU (EPTUNU)

Universitas Nahdlatul Ulama (UNU) Purwokerto has established the Unit Pelaksana Teknis (UPT) Bahasa as a university-level language center responsible for managing and improving students' English proficiency. The establishment of UPT Bahasa supports the university's efforts to enhance students' communicative competence and academic readiness. University language centers commonly function as units that provide language instruction, testing, and certification to meet academic and institutional needs (Nurhayati, 2021).

One of the main responsibilities of UPT Bahasa is administering the English Proficiency Test of UNU Purwokerto (EPTUNU). EPTUNU is an institutional English proficiency test designed to assess students' abilities in Listening Comprehension, Structure and Written Expression, and Reading Comprehension. The implementation of EPTUNU reflects the university's policy to ensure that students achieve a minimum level of English proficiency before completing their academic programs. Similar institutional English proficiency tests are widely implemented in Indonesian universities as part of graduation requirements (Musahadah, 2015).

UPT Bahasa is responsible for preparing, administering, and scoring EPTUNU. In addition to test administration, the unit also offers English preparation classes and training programs to help students meet the required proficiency level. The test results are used as academic requirements for thesis proposal submission, thesis examination, and graduation. In this role, UPT Bahasa functions not only as a testing unit but also as an academic support unit that contributes to students' language development (Rawlusyk, 2018).

The administration of EPTUNU follows standardized procedures to ensure fairness and reliability. Students register for the test through the UPT

Bahasa system and take the test according to scheduled sessions held each semester. During the examination, students are required to follow established test regulations. The test is scored internally by UPT Bahasa staff, and the results are announced after the completion of the scoring process. Students who meet the required score receive an official EPTUNU certificate as proof of fulfilling the university's English proficiency requirement.

In developing EPTUNU test materials, UPT Bahasa adapts various resources from external references and practice materials. The test format follows a TOEFL-like model, consisting of Listening Comprehension, Structure and Written Expression, and Reading Comprehension. This format aligns with commonly accepted English proficiency testing models used in higher education institutions (Fatimatazzahro, 2022). The adaptation of materials is conducted to ensure that the test meets institutional objectives while maintaining content validity and practicality (Fulcher, 2015).

Overall, EPTUNU plays an important role in supporting academic quality at UNU Purwokerto. The test ensures that graduates possess adequate English proficiency as part of the university's academic standards. EPTUNU results are used for academic and administrative purposes, particularly as requirements for thesis completion and graduation. Through the implementation of EPTUNU, UNU Purwokerto has established a structured system of English language assessment that supports student achievement and institutional development.

D. Competences in English Proficiency Test

The English Proficiency Test (EPT) measures students' ability to understand and use English in academic and professional situations. According to Fulcher (2015), language proficiency tests evaluate a person's performance in several main language skills, such as listening, reading, speaking and writing. In many universities, especially in Indonesia, EPTs mainly focus on receptive skills, including listening comprehension, structure and written expression, and reading comprehension (Widodo & Renandya, 2016). These skills are important because they show how well students can understand spoken and written

English. Aryadoust (2019) also explained that assessing these skills help ensure the test accurately reflects the learners' true English ability. Therefore, the EPT not only measures students' language skills but also helps determine their readiness to study and communicate in English-based academic environments.

1. Listening Comprehension

The first section in this test is listening comprehension. Listening itself is a receptive skill. Listening comprehension is designed to assess the ability of test takers in understanding spoken English. According to (Brown, 2004) listening comprehension is a person's ability to understand what they heard from native speakers. In this section, the students as test takers are expected to know the meaning obtained through contextual information based on the knowledge they learned. Test takers must listen to several types of recorded passages and answer multiple choice questions on them (Phillips, 2011). Listening carefully is important because the recording is only played once and the material is not stated in the test book.

The listening sections comprises 3 different parts when the test takers must answer 50 questions in 30-40 minutes The first is Part A consists of 30 questions which test takers will hear a short conversation between 2 speakers that ends with a question. The next is Part B listening to long conversations up to two minutes by the speakers with 8 number of questions. The last is Part C with 12 questions contain the lectures or talks which include factual topic conveyed by the speaker with long durations.

2. Structure and Written Expression

Structure & Written expression come next. This section assesses the ability of comprehending formal written English since many things that are acceptable in spoken English may not be accepted formally in English. It aims to evaluate the test takers' ability in recognizing English sentences with grammatically correct. They must identify the grammatical error in sentences also choose the correct answers to complete sentences.

This section comprises of two parts which are (A): Structure and (B): Written expression with the total of 40 questions where the takers have 25 minutes to answer them. Part A is structure contains that has 15 multiple-choice questions with A, B, C, and D probable choices for sentence completion, whereas Part B contains 25 questions. In the second phase, test takers will experience type of error-analysis questions in each sentence.

3. Reading Comprehension

The final section of the English Proficiency Test is Reading Comprehension, which plays a significant role in evaluating students' ability to process written information. According to Aida (2021), reading is a method of gaining knowledge and information through written text. Similarly, Nuttall (2002) emphasized that reading involves an interactive process between the writer's and the reader's minds, whereby readers interpret the writer's intended meaning to grasp the underlying message. From these perspectives, reading can be defined as the process of deriving meaning from text in order to obtain knowledge, information, or messages depending on the writer's communicative purpose.

Reading, therefore, holds an essential position in education, as it is one of the primary means through which learners acquire knowledge. As noted by Suyanto et al. (2024), the skill of reading is crucial for educational success, given that access to information largely depends on one's ability to comprehend written materials. In the context of proficiency testing, reading comprehension extends beyond simple decoding of text; it assesses an individual's ability to comprehend, identify, and critically analyze various types of reading passages. This makes it a reliable measure of test takers' ability to understand written English at different levels of complexity.

In the Reading Comprehension section of the test, students are presented with passages followed by multiple-choice questions. These questions are designed to evaluate the ability to grasp main ideas, identify supporting details, infer implied meanings, and understand the meaning of specific words or phrases within context. As English vocabulary often

contains words with multiple meanings, this section particularly emphasizes contextual understanding (Suyanto et al., 2024). Thus, test takers must interpret vocabulary and expressions in relation to the passage rather than relying on dictionary definitions.

The passages used in this section generally cover academic topics and are written in styles similar to those encountered in higher education contexts. Test takers are expected to read a variety of short passages and respond to several questions for each passage. The questions target both explicitly stated information and implicit ideas, as well as the correct interpretation of specific lexical items. In this way, the Reading Comprehension section provides a comprehensive measure of students' ability to process, interpret, and analyze written English, making it a vital component of the overall English Proficiency Test.

E. Approaches in Measurement

In the field of educational and psychological testing, there are two primary approaches to measurement: the classical approach and the modern approach, commonly known as Classical Test Theory (CTT) and Item Response Theory (IRT). Both frameworks aim to ensure the validity, reliability, and precision of test instruments; however, they differ in their conceptual foundations, analytical methods, and assumptions about measurement error.

1. Classical Test Theory (CTT)

Classical Test Theory (CTT) represents the traditional foundation of test development and analysis in educational and psychological measurement. According to Allen & Yen (2002), CTT assumes that every observed score obtained by a test taker consists of two components: a true score, which reflects the examinee's actual level of ability or knowledge, and an error score, which accounts for random factors that may influence test performance. This model provides a conceptual basis for understanding how test scores represent both the learner's true ability and the measurement error inherent in testing.

CTT employs several important statistical indices to evaluate test quality, including reliability coefficients, item difficulty, and item discrimination. Reliability coefficients measure the internal consistency or stability of a test, while item difficulty and discrimination indices assess how well individual items function in distinguishing between different levels of ability among test takers. These analytical tools make CTT a valuable framework for evaluating the quality and fairness of test items.

In the context of educational measurement, Prasetyo (2018) explains that classical item analysis involves examining students' responses to evaluate and improve test quality. Each item is analyzed in terms of its level of difficulty, discrimination power, and distribution of answer choices (particularly for multiple-choice items). This process helps test developers identify which items perform effectively and which may require revision or replacement.

The advantages of Classical Test Theory lie in its practicality and accessibility. As noted by Prasetyo (2018), CTT is cost-efficient, simple to implement, and can be easily applied using computer-based tools. It is also suitable for use with small samples and provides immediate feedback for item improvement, making it widely used in schools and educational institutions for everyday test analysis.

However, CTT is not without limitations. One of its major drawbacks, as described by both Allen and Yen (2002) and Mardapi (2012), is that its results are sample-dependent that is, item statistics such as difficulty and discrimination may vary depending on the group of examinees analyzed. Additionally, the estimation of a test taker's ability depends on the specific set of items administered, limiting generalizability. The standard error of measurement is assumed to be constant for all examinees, which may not accurately reflect individual variability. CTT also provides limited insight into the interaction between individual test takers and specific items. Moreover, reliability estimation often relies on the assumption of parallel or equivalent test forms, which is not always practical.

Despite these limitations, Classical Test Theory remains one of the most widely applied models in educational and psychological testing due to its conceptual clarity, computational simplicity, and long-standing role in measurement theory. Its contributions continue to provide a foundation for understanding test reliability and validity, while serving as a basis for comparison with more advanced frameworks such as Item Response Theory (IRT).

2. Modern Test Theory (IRT)

Modern Test Theory, commonly referred to as the Item Response Theory (IRT) or Latent Trait Theory (LTT), represents a significant advancement over the traditional framework of Classical Test Theory. According to Prasetyo (2018), modern item analysis is a process of evaluating test items using mathematical models that describe the probabilistic relationship between a test taker's ability and the likelihood of answering an item correctly. This approach recognizes that a learner's response to a test item is influenced not only by their ability level but also by the characteristics of the item itself. The graphical representation of this relationship is known as the Item Characteristic Curve (ICC), which illustrates how the probability of a correct response changes as the examinee's ability increases.

Item Response Theory (IRT) provides several key advantages that distinguish it from Classical Test Theory. As outlined by Prasetyo (2018), IRT (1) produces sample-independent item parameters, meaning item statistics remain stable across different examinee groups; (2) yields test-independent person parameters, allowing ability estimates to remain constant regardless of the specific test form administered; (3) focuses on item-level analysis rather than test-level statistics, providing more detailed insights; (4) eliminates the need for parallel tests to estimate reliability; and (5) allows for precision measurement at different levels of ability, as each ability estimate is accompanied by its own standard error. These advantages

make IRT a powerful framework for ensuring objectivity, fairness, and precision in modern measurement.

According to Hambleton et al. (2011), several IRT models have been developed, each differing in the number and nature of parameters estimated:

1) One-Parameter Model (Rasch Model) focuses solely on the item difficulty parameter. It assumes all items have equal discrimination power and no guessing effect.
2) Two-Parameter Model (2PL) considers both item difficulty and item discrimination, allowing for variation in how well each item distinguishes between high- and low-ability examinees.
3) Three-Parameter Model (3PL) expands the model by adding a guessing parameter, which accounts for the probability that an examinee with low ability may still answer correctly by chance.
4) Four-Parameter Model (4PL) further extends the 3PL model by including an additional parameter that represents other external factors influencing the probability of a correct response, such as test-taking motivation or item irregularities.

Beyond these model variations, IRT represents a broader psychometric paradigm that addresses many of the limitations found in Classical Test Theory. As described by Embretson and Reise (2013), IRT mathematically models the probability of a correct response as a function of an examinee's latent ability. The model produces sample-independent and test-independent estimates, meaning both item and person parameters remain stable across different groups and test forms. Moreover, IRT provides detailed item-level diagnostics, including estimates of item difficulty, discrimination, and guessing parameters, allowing for a deeper understanding of how individual items perform within a test.

IRT's strengths have made it especially valuable in large-scale assessments, computerized adaptive testing (CAT), and test validation studies. Its ability to provide precise measurement across a wide range of ability levels ensures fairness and accuracy in evaluation. When

implemented through specialized software such as QUEST, IRT enables researchers and educators to obtain precise estimates of test-taker ability, improve item calibration, and construct assessments that are both reliable and equitable. By combining mathematical rigor with practical application, IRT has become the foundation of modern measurement and continues to shape contemporary approaches to educational and psychological testing.

F. Item Analysis

Item Analysis is related to the several items of statistical analysis in analyzing characteristics and features of a test. They consist of validity, reliability, level of difficulty, and discriminating power.

1. Validity

Validity refers to the extent to which a test accurately measures what it is intended to measure and the degree to which the interpretations and decisions based on test scores are appropriate and supported by empirical evidence. According to Bachman (2004), validity is an integrated evaluative judgment concerning the adequacy and appropriateness of inferences and actions derived from test results. Similarly, Wahidmurni (2014) emphasizes that validity must be supported by empirical evidence demonstrating that the interpretation of test scores aligns with the intended construct being measured.

To establish validity, experts must examine whether the test content accurately represents the skills and objectives it aims to assess. This involves analyzing the consistency between the syllabus, test objectives, and the actual test items. Brown (2016) explains that if the test items adequately reflect the objectives derived from the syllabus, the test can be said to possess content validity. This view is supported by Hughes (2005), who defines content validity as the extent to which a test's content represents a comprehensive sample of the language skills, structures, and competencies it is designed to assess. In other words, a test achieves content validity when its items appropriately reflect the intended learning outcomes and skills.

Within the framework of Item Response Theory (IRT), validity is evaluated not through a single correlation coefficient but through the extent to which the statistical model fits the observed data. IRT conceptualizes the relationship between test-taker ability (θ) and item characteristics such as difficulty and discrimination using probabilistic mathematical models. A test is considered valid in the IRT framework when the model accurately predicts examinees' response patterns and when item parameters represent meaningful distinctions in ability levels. The fundamental IRT model for dichotomous data (1-parameter logistic or Rasch model) is expressed as:

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}$$

Where:

$P_i(\theta)$ = probability of a correct response to item i

θ = test taker's latent ability

b_i = item difficulty parameter

Validity in IRT is demonstrated when the observed response patterns of test takers fit the expected model. This is evaluated through fit statistics, such as Infit Mean Square (MNSQ) and Outfit Mean Square (MNSQ) values (Bond & Fox, 2015):

- Ideal values are approximately 1.0 (acceptable range: 0.7–1.3).
- Values significantly above or below this range indicate misfit (invalidity at item or person level).

Additionally, construct validity in Item Response Theory (IRT) can be examined using the Wright Map (also known as the Person–Item Map). This map visually represents the alignment between test-takers' ability levels and the difficulty of each test item. A well-constructed test will show that items are appropriately distributed across the ability range, ensuring that both high- and low-ability students are accurately measured.

Furthermore, in Item Response Theory (IRT), test validity is reinforced through its ability to examine how well each item functions across varying levels of examinee ability. IRT evaluates the probabilistic relationship between a test-taker's latent ability and their responses, allowing researchers to determine whether items consistently measure the intended construct. This approach provides detailed diagnostic information at the item level, enabling a more precise evaluation of item quality, construct representation, and overall test structure. By focusing on item functioning rather than relying solely on total test scores, IRT offers a more comprehensive and informative framework for establishing test validity, ensuring that assessment results are both meaningful and aligned with the construct being measured. IRT uses Mean Square (MNSQ) fit indices to check whether items behave as expected.

Table 1.1. Validity Range Interpretation

MNSQ Range	Interpretation
0.7 – 1.3	Ideal fit (commonly accepted for educational tests)
0.5 – 1.5	Acceptable for most testing situations
< 0.5	Overfit (item too predictable, not contributing new information)
> 1.5	Misfit (item not measuring the intended construct)

2. Reliability

Reliability refers to the consistency of test result. Reliable here means that a test must rely and fit on several aspects in conducting the test itself. A test should be reliable toward students. Bachman (2004: 153) states that reliability is consistency of measures across different conditions in the measurement procedures. Test administration must be consistent by which a test can be said as well-organized test. In vice versa, bad administration

and unplanned arrangements of a test can make it does not work in measuring students accomplishment.

An evaluation instrument is considered to have high reliability when the test produces consistent results in measuring what it aims to measure. Reliability ensures consistency, which fulfills the essential requirement that the test scores are valid. The more reliable a test is, the more confident we can be that it will yield the same results when administered again.

In IRT, reliability is closely associated with the Test Information Function (TIF), which indicates the amount of statistical information or precision obtained from the test at various ability points. The relationship between reliability is expressed through the following formula:

$$\text{Reliability}(\theta) = \frac{I(\theta)}{I(\theta) + 1}$$

Where:

I(θ) = test information at ability level θ

Higher information values indicate higher measurement precision (and thus reliability) at that level of ability (Embretson & Reise, 2013; Bond & Fox, 2015). This formula shows that as the amount of information increases, the reliability of the test at that ability level also increases. Therefore, a well-constructed test under IRT provides high information (and thus high reliability) across the range of abilities it is intended to measure.

Although IRT does not prescribe a fixed "one-value" reliability coefficient, the following general guidelines are widely accepted for interpreting IRT-based reliability across the ability scale:

Table 1.2. Reliability Range Interpretation

Approximate Reliability	Interpretation
> 0.90	Excellent precision; ideal for high-stakes tests (e.g., proficiency exams, certification tests).
0.80 – 0.89	Good reliability; acceptable for most educational assessments, including language tests
0.70 – 0.79	Adequate; suitable for low-stakes tests or early-stage assessments.
< 0.70	Weak precision; indicates that items do not provide enough information to reliably measure ability.

3. Level of Difficulty

A good test is one that maintains an appropriate balance between difficulty and ease for the test-takers. It should not be overly simple or excessively challenging, but instead provide a fair measure of students' abilities. The test items should include reasonable answer options that are close to the correct answer, encouraging students to think critically. As Brown (2004) explains, very easy items can help build a sense of success among lower-ability students and serve as warm-up questions, while very difficult items can offer a challenge to higher-ability students. This balance ensures that tests can effectively assess a range of abilities and motivate students. If a test is consistently too easy or too difficult, it fails to provide accurate information about students' proficiency levels and the teacher's assessment standards. Therefore, a well-constructed test must meet the characteristics of a good test and adhere to appropriate standards of measurement.

The degree of test difficulty is represented by a numerical value known as the difficulty index (Arikunto, 2006). This index has both

minimum and maximum limits; the lower the index value, the more difficult the test item, and conversely, the higher the index value, the easier the item. According to Mehrens & Lehmann (1984), several factors must be considered when determining the appropriate level of item difficulty, including (1) the purpose of the test, (2) the ability level of the students, and (3) the age or grade level of the examinees.

Item Response Theory (IRT) conceptualizes item difficulty not as a proportion but as a *parameter* that describes the point on the ability scale (θ) where the probability of a correct response is 50%. This relationship is defined mathematically by the logistic function, as follows (Embretson & Reise, 2013):

$$P_i(\theta) = \frac{1}{1 + e^{-Da_i(\theta - b_i)}}$$

where:

$Pi(\theta)$ = probability that a person with ability θ answers item I correctly,

a_i = item discrimination parameter,

b_i = item difficulty parameter,

D = scaling constant (usually 1.7),

e = base of the natural logarithm.

Ideal range: $-2.0 \leq b \leq +2.0$ (Embretson & Reise, 2013). The IRT model provides a robust and sample-independent estimation of item difficulty, allowing test items to be evaluated consistently regardless of the specific group of examinees. This characteristic makes IRT particularly suitable for standardized language assessments such as the English Proficiency Test (EPT) or TOEFL, where stable and precise measurement across different test populations is essential.

Table 1.3. Item Difficulty Range Interpretation

***b*-value**	**Interpretation**
–3.0 to –2.0	Very easy items (Need review)
–2.0 to –1.0	Easy items (Acceptable)
–1.0 to +1.0	Moderate (Ideal for most tests)
+1.0 to +2.0	Hard items (Acceptable)
+2.0 to +3.0	Very hard items (Need review)

4. Dicrimination Power

It is the extent to which an item differentiates between high and low-ability test-takers. Discrimination is important because if the test-items can discriminate more, they will be more reliable (Hughes, 2005:226). It can be defined also as the ability of a test to separate master students and non-master students (Arikunto, 2006). A master student is a student with higher scores of test, and a non-master student is a student with lower scores on the test given. The same as the term of difficulty level, discrimination has discrimination index. It is an indicator of how well an item discriminates between weak candidates and strong candidates (Hughes, 2005:226). This index is used to measure to the ability of a test in discriminating the upper and lower group of students. Upper students are students who answer with true answer, and lower group are students with false answer. In this index, it has negative point. Different from difficulty index, the negative index of discrimination power shows that the questions identify high group students as poor students and low group students as smart students. A good question is a question that can be answered by upper group and cannot be answered with true answer by lower group.

In Item Response Theory (IRT), item discrimination is represented by the parameter *a* in the logistic function, which reflects how sharply an item distinguishes between examinees with different ability levels. Items with high discrimination values can clearly separate high- and low-ability

test takers, while those with low discrimination are less effective in doing so. In the context of an English Proficiency Test (EPT), good item discrimination ensures that each test item accurately measures differences in students' language abilities, leading to a fair and reliable assessment. The formula for item discrimination in IRT can be expressed as:

$$P(\theta) = \frac{1}{1 + e^{-a(\theta - b)}}$$

Where:

a = item discrimination parameter

b = item difficulty parameter

θ = examinee's latent ability level

A higher a-value indicates that the item more effectively differentiates between examinees near the item's difficulty level (b). In contrast, a low a-value suggests the item does not distinguish well between high- and low-ability test-takers.

Table 1.4. Item Discrimination Range Interpretation

Infit MNSQ	**Interpretation**	**Quality Judgment**
> 1.21	Weak discrimination	Needs review
0.71 – 1.20	Acceptable discrimination	Good
0.80 – 1.10	Ideal discrimination	Excellent
0.51 – 0.70	Very strong discrimination	Overfit

G. Quest Program

QUEST is a specialized data analysis software widely recognized for its advanced capabilities in questionnaire and test analysis, particularly through the application of Rasch measurement theory. The program enables the construction and validation of variables based on both dichotomous and polytomous observations, such as multiple-choice responses, Likert-type rating scales, and partial-credit items. By integrating the most recent developments in psychometric theory with user-friendly functions, QUEST has become an important tool in the fields of educational and language assessment.

The QUEST (Question Analysis and Test Scoring) program was developed by Adams and Khoo (1996) as an instrument for conducting Item Response Theory (IRT)-based psychometric analysis. Unlike Classical Test Theory (CTT), which primarily relies on raw scores and sample-dependent statistics, QUEST employs Rasch modeling to generate more precise and sample-independent measurements (Bond & Fox, 2015). This methodological advantage allows researchers and educators to evaluate key test properties such as item difficulty, item discrimination, reliability, and model fit, thereby ensuring that assessments meet rigorous standards of validity and fairness.

Through Rasch analysis, QUEST produces detailed estimates of both item characteristics and respondent abilities, along with diagnostic fit statistics. The program's output includes item-level indices such as counts, percentages, and point-biserial correlations, as well as person-level measures of ability. These results can be presented in comprehensive tables or visual maps, allowing researchers to interpret patterns of performance and identify problematic items with greater clarity.

Adams and Khoo (1996) highlight QUEST as a comprehensive analysis environment that combines modern Rasch measurement theory with traditional statistical procedures. Its flexible command-based system supports a wide variety of test formats, including multiple-choice items, rating scales, short-answer items, and completion-type questions. Moreover, QUEST features an

accessible control language and generates flexible, informative outputs, which makes it suitable for both research and applied testing contexts.

A notable strength of QUEST lies in its ability to accommodate different analytical modes. The batch mode allows for efficient processing of large datasets, making it ideal for standardized testing environments. The interactive mode, facilitated through the display manager, provides researchers with real-time data exploration. Additionally, a hybrid mode combines these two approaches, offering flexibility in analysis. By calibrating both test-taker ability and item difficulty on a shared logit scale, QUEST represents results along a unidimensional axis in log-odds units, enabling direct comparison between examinee performance and item characteristics. This ensures objective, fair, and accurate measurement in test construction.

The strengths of QUEST include its measurement precision, operational flexibility, and analytical efficiency. It streamlines the psychometric analysis process while maintaining strict accuracy standards, making it particularly valuable in both academic research and practical assessment contexts. The program's capacity to handle multiple item types, deliver detailed diagnostic statistics, and generate graphical representations positions it as an indispensable tool for the development, evaluation, and refinement of educational and psychological assessments.

The advantages of the Quest program include:

1) Quest can perform analyses to estimate both item groups and respondent groups.
2) Quest provides not only item and respondent estimates but also displays the relationships between them.
3) Quest allows for correction of erroneous item or respondent data through a simplified procedure.
4) Analysis results from Quest can be transferred to and opened with various programs, such as Microsoft Office, Notepad, and WordPad.
5) Quest can display and record scores for each item and each respondent.
6) Quest can analyze data with varying scoring criteria for each item.

Operationally, QUEST functions as an interactive, command-driven system. Users enter data and commands directly into the program's interface (syntax screen). For successful operation, both the data file and the program file must be stored in the same directory. Analysis results are then produced as output files, which can be accessed and further examined through applications such as Microsoft Word, Notepad, or WordPad. Command statements in QUEST generally consist of four parts: the command word, arguments, options, and redirection (Wu et al., 2007). This structured system provides both flexibility and precision, enabling users to tailor their analyses according to specific research needs.