

# PSYCHOMETRIC ANALYSIS OF A TOURISM ENGLISH TEST IN INDONESIA: A CLASSICAL TEST THEORY APPROACH

Rezkilaturahmi<sup>1</sup>, Heri Retnawati<sup>2</sup>, Nur Hidayanto Pancoro Setyo Putro<sup>3</sup>, La Ode Ilan<sup>4</sup>, Ilyas<sup>5</sup> and Taufik Rizki Sista<sup>6</sup>

<sup>1</sup>Universitas Negeri Yogyakarta, Educational Research and Evaluation

<sup>2</sup>Universitas Negeri Yogyakarta, Educational Research and Evaluation

<sup>3</sup>Universitas Negeri Yogyakarta, English Language Education

<sup>4</sup>Universitas Negeri Yogyakarta, English Language Education

<sup>5</sup>Universitas Negeri Yogyakarta, Educational Research and Evaluation

<sup>6</sup>Universitas Darussalam Gontor, Islamic Education

<sup>1</sup>rezkilaturahmi.2023@student.uny.ac.id, <sup>2</sup>heri\_retnawati@uny.ac.id,

<sup>3</sup>nur\_hidayanto@uny.ac.id,

<sup>4</sup>laodeilan98@gmail.com, <sup>5</sup>ilyas.2023@student.uny.ac.id and

<sup>6</sup>taufikrizki90@unida.gontor.ac.id

<sup>1</sup>Orcid ID: <https://orcid.org/0009-0003-8449-6970>, <sup>2</sup>Orcid ID <https://orcid.org/0000-0002-1792-5873> and <sup>3</sup>Orcid ID: <https://orcid.org/0000-0002-3019-8498>

## Abstract

Assessment plays a strategic role in ensuring that English learning for vocational tourism students aligns with industry and cultural demands. However, limited research has examined the psychometric quality of tourism English assessments integrated with cultural insight. This study aims to analyze the psychometric properties of a Tourism English Test based on the Cultural Insight in English (CIE) framework using a Classical Test Theory (CTT) approach. The study employed a quantitative evaluative design involving 90 vocational high school students majoring in tourism in Indonesia. Content validity was examined using Aiken's V through seven expert judgments, followed by construct validation using Confirmatory Factor Analysis (CFA). Item characteristics were further analyzed using CTT parameters, including item difficulty, discrimination index, and reliability estimation. The results indicated that 22 items met acceptable content validity criteria after expert review. CFA supported a dominant unidimensional structure with Reading and Grammar as core components of CIE ability. CTT analysis revealed that most items demonstrated moderate difficulty levels and good discrimination indices, although several items required revision due to low or negative discrimination values. The instrument showed good internal consistency reliability (Cronbach's Alpha = 0.797). Overall, the findings confirm that the Tourism English Test based on the CIE framework possesses adequate validity and reliability and is suitable for measuring culturally embedded English proficiency among vocational tourism students. The study contributes to strengthening assessment practices in vocational English education by integrating cultural insight into psychometrically sound measurement models.

**Keywords:** English Test, Assessment, Tourism, Classical Test Theory, Validity, Reliability

## 1. INTRODUCTION

Assessment in English language learning for vocational students is important to assess how far students have understood the concepts and language skills related to the world of work. According to (Purwati et al., 2023) emphasize that assessment in this context is not only to measure students' general language abilities, but also as a tool for teachers to evaluate the effectiveness of learning methods. Similarly as Matondang et al., (2022) and Tsai (2024) With input obtained through assessment, teachers can optimize their approach, for example by emphasizing more on aspects of language that are specific to students' vocations, so that learning is more relevant and applicable to professional needs.

In the context of vocational high school, the ideal assessment must not only measure technical skills, but also the ability of students to understand and apply cultural values (Handayani et al., 2018; Pambayun

PSYCHOMETRIC ANALYSIS OF A TOURISM  
ENGLISH TEST IN INDONESIA: A CLASSICAL  
TEST THEORY APPROACH

et al., 2020, 2023). The context of cultural insight in assessment emphasizes evaluations that include students' understanding of local cultural values and their application in their field of expertise. This is in line with the integration of local wisdom in education can enrich students' perceptions of their own culture and support sustainable development (Suaco et al., 2023). In other words, the assessment at vocational high school does not only function as a measure of learning outcomes, but also to strengthen the relationship between vocational skills and local wisdom in education.

The four language skills play a significant role in helping them understand material that is directly related to the industrial world (Uktolseja & Gaspersz, 2019). It means that integrating various cultural elements into educational materials into good methods because it will be motivated the students' learning (Hulu et al., 2024). This will be very better if the assessment used is also based on cultural elements. In addition, according to (Garing et al., 2023) asserted that education becomes more interesting when teachers use materials that are rooted in cultural elements. This approach increases students' motivation and enthusiasm because the material resonates with their environment. In this case, the assessment used should also be backed by culture. However, the assessment of various types of models used in several vocational school schools has not been cultural -based and has not yet been analysed in depth, both small and broad scale tests with the ADDIE development model. Similarly as Dieu & Dan (2023) and Nggawu & Thao (2023) suggests that English language learning for tourism students' needs to be designed by integrating various language learning strategies, such as memory (memorization), cognitive, metacognitive, compensatory, affective, and social strategies. These strategies are considered relevant because tourism students are required not only to master linguistic aspects but also to be able to use English flexibly in real-life communication situations. The study confirmed that students with a tourism background face unique challenges, particularly when interacting directly with international tourists. These challenges include limited ideas, lack of confidence, and difficulty adapting the language to diverse cultural contexts. This situation suggests that tourism English learning requires a more contextual and applied approach so that students can respond spontaneously and appropriately to communication situations.

In line with this, (Novita et al., 2022) analyzed the online learning process in vocational education, specifically for tourism students. The results showed that one of the main obstacles in online learning is the limited knowledge and skills of educators in developing and utilizing learning media appropriate to the characteristics of vocational education. In fact, English language learning for tourism students ideally emphasizes direct communication practice over theoretical delivery. The lack of practical activities results in students lacking practical training in using English functionally. This finding reinforces (Dieu & Dan, 2023) view that tourism students often experience a lack of ideas and difficulty acting when confronted with tourists in real-life situations.

Various previous studies have shown that English language learning for vocational high school (VHS) students, particularly those in the tourism specialty, is a particular concern. These studies emphasize the importance of language learning strategies (Dieu & Dan, 2023; Evenddy et al., 2022; Islam et al., 2025; Kotollaku, 2024; Novita et al., 2022; Thi & Trang, 2023) strengthening communication practices, developing materials based on the needs of the tourism industry, integrating cross-cultural competencies, and utilizing innovative technology and media such as virtual tours. Furthermore, several studies have highlighted the role of English in supporting the tourism sector at a macro level, both from an economic perspective and in the readiness of professional and competent tourism human resources at the ASEAN regional level.

However, the majority of these studies still focus on the learning and material development aspects, while assessment, a key component of the English language learning system, has not received adequate attention. In the context of vocational education, assessment plays a strategic role not only in measuring learning outcomes but also in ensuring the alignment of graduates' competencies with the needs of the workforce and the tourism industry. In other words, the success of tourism English learning cannot be separated from the quality of the assessment model used.

Furthermore, previous studies tended to view English language assessment conventionally, limited to measuring linguistic aspects such as grammar, vocabulary, and reading comprehension (Pelzer, 2019; Tambunan & Kurniawan, 2024). This type of assessment does not fully reflect the complexity of competencies required by tourism vocational high school students, which require the ability to

communicate functionally, contextually, and sensitively to cultural differences. Thus, there is a clear gap between the communication competency demands in the tourism industry and the English language assessment models currently used in vocational high schools.

On the other hand, several studies have emphasized the importance of cross-cultural competence in tourism English learning. However, the integration of these cultural aspects has largely remained at the level of teaching materials and learning activities, and has not yet developed into a systematic, structured, and empirically tested assessment model. As a result, teachers lack assessment instruments capable of measuring the extent to which students truly understand and are able to apply cultural insights to English communication in the tourism context. This indicates a significant research gap between the concept of culture in learning and its implementation in the assessment system.

## 2. PROBLEM STATEMENT

Furthermore, although there is research highlighting the implementation of the ASEAN tourism curriculum and regional competency standards (Waris, 2025), these studies have not specifically developed English language assessment models aligned with the local context, particularly the regional culture that is a major tourist attraction. In the context of Yogyakarta, known as a leading cultural tourism destination, the richness of local values, traditions, and wisdom should be the primary source in tourism English learning and assessment. However, to date, there has been little research explicitly integrating Yogyakarta's local culture into an English language assessment model based on Cultural Insight in English (CIE).

Another gap is evident in the separation between learning models and assessment models. Previous research generally discussed the two separately, thus failing to produce a comprehensive and integrated framework. In fact, in modern learning approaches, particularly in vocational education, assessment should be an integral part of the learning process, not merely a final stage to determine learning outcomes. Without an assessment model aligned with a culture-based learning approach, the learning process has the potential to become disoriented and unable to provide meaningful feedback for both students and teachers. Therefore, this research is highly urgent because it attempts to fill this gap by developing an English language proficiency assessment model based on Cultural Insight in English (CIE), specifically designed for vocational high school students majoring in tourism in Yogyakarta. This assessment model serves not only as an assessment tool but also as an integral part of a culture-based English language learning model, emphasizing the integration of language competency, cultural understanding, and the needs of the tourism industry. Thus, assessment is no longer merely administrative but becomes a pedagogical tool to encourage more meaningful and contextual learning. In addition, the development of the CIE-based assessment model also contributes novelty from a methodological perspective, because it combines a development approach (ADDIE), an innovative question format (Multiple-Choice Multiple-Answer), and comprehensive instrument analysis through Classical Test Theory (CTT) and Item Response Theory (IRT). This approach has not been widely used in tourism English assessment research, especially at the vocational high school level. Therefore, although various previous studies have emphasized the importance of tourism English learning, cross-cultural competence, and student readiness to face the global workforce, there is still a significant gap in terms of developing an integrated, contextual, and local culture-based assessment model. Therefore, this research is very important and relevant to conduct, because it offers a comprehensive solution through the development of an assessment model while strengthening the English learning model based on Cultural Insight in English for vocational high school students specializing in tourism in Yogyakarta. The research aimed to investigate the quality of the English language ability within the Cultural Insight in English (CIE) for vocational high school students majoring in tourism in Yogyakarta. The qualities include the content validity, the construct validity and the item characteristics of the English language ability assessment model within the Cultural Insight in English (CIE) for vocational high school students majoring in tourism in Yogyakarta.

## 3. THEORETICAL FRAMEWORK

A good assessment process will determine good learning outcomes because through assessment it can

PSYCHOMETRIC ANALYSIS OF A TOURISM  
ENGLISH TEST IN INDONESIA: A CLASSICAL  
TEST THEORY APPROACH

also be seen to what extent students have mastered their learning, then assessment can also provide an overview in assessing the level of curriculum achievement (Lipnevich et al., 2013). Thus, assessment plays a crucial role in education by evaluating student mastery, guiding curriculum achievement, and ensuring the quality of learning outcomes.

Learning assessment has systematic rules and steps. Unfortunately, specifically T. C. Brown & Peterson (2009) states that one understanding between assessment and testing often occurs especially in the field of education. However, basically, tests have administrative procedures and are carried out at the time specified in the curriculum with a systematic performance measurement and evaluation framework. Meanwhile, Kleij (2019) assessment itself is an ongoing process that covers a wider domain such as when students provide comments, respond to an explanation in the assessment process and provide ideas are part of the assessment process. Thus, tests are one of the tools used by educators/teachers to assess the extent of a student's abilities.

Assessment requires the use of certain techniques to measure student achievement but in a systematic process that starts from the stage of identifying learning objectives, monitoring student progress towards the learning objectives that have been set to the extent to which learning objectives have been achieved (Scholkmann et al., 2023). Furthermore, the requirements in the assessment include all methods that can be used to measure the success of student learning in the learning process that is assessed individually or in groups.

Assessment is carried out continuously as a goal to be able to improve the quality of learning to the maximum. As a process that is carried out continuously, Cooper et al. (1988) stated that "assessment is the process of collecting, synthesizing and interpreting information to assist in decision making". The objectives of conducting an assessment include: (a) building a class that supports learning so that the class becomes a positive learning environment, (b) through assessment, student abilities can be identified so that they can help make decisions about the placement of good student success, (c) providing feedback to help improve student learning and behaviour, (d) diagnosing student problems and difficulties in learning, and (e) summarizing and assessing learning.

The complex assessment process makes it a systematic and long step. Cooper et al. (1988) continued that because assessment is complex and is *"a process for documenting, in measurable terms, the knowledge, skills, attitudes, and beliefs of the learner"*, therefore teachers as assessors of a learning process need to specifically determine which knowledge, skills, attitudes and beliefs require assessment in what part and for what purpose should be assessed and what assessment tools are good and appropriate to complete the assessment process. This is in line with how Ghouali & Bekr (2021) stated that assessment is a series of activities to obtain, analyze, and interpret data about the process and results of student learning which are carried out systematically and continuously so that they become meaningful information in making decisions about student learning outcomes. Furthermore, Arifin (2017) also stated that assessment is a process or activity that is carried out systematically and continuously to collect information about the process and results of student learning to make decisions based on criteria and predetermined standards. Previous definitions of assessment are in line with Setiawati et al. (2013) that assessment has a specific focus, namely the learning achievements that have been achieved by students so that the assessment process aims to obtain evidence related to student learning achievements.

Another opinion was put forward by Hung et al. (2012) that *"Assessment goes beyond just giving a test and scoring it, in fact, assessment is intended to provide both students and teachers with a detailed view of the achievement toward the course goals within the teaching and learning process"*. An assessment activity is more than just giving tests and scoring because in fact the assessment is intended to provide students and teachers with a detailed view of the achievement of learning objectives in the teaching and learning process. In this case, an assessment activity directly provides benefits for teachers and students at the same time.

The Indonesian Education System has been completely described in Permendikbud No. 23 of (2016) concerning Basic and Secondary Education Assessment Standards which states that assessment is the process of collecting and processing information to measure student learning outcomes. This process is carried out through various assessment techniques, using various instruments and coming from various sources to be more comprehensive. Assessment must be carried out effectively so that the collection of

information that will be used to measure student learning achievement must be complete and accurate in order to produce the right decision (Ajayi & Idusogie, 2020). Based on the series of definitions above, it is concluded that assessment and learning have a very close relationship and cannot be separated. Assessment is a structured activity that often takes place both at the beginning of learning, in the learning process or at the end of learning. The results obtained from the next assessment will be a means to improve and enhance the quality of learning itself. This is because through assessment it will be known to what extent students have mastered learning within a certain period. Thus, results of the assessment also allow detection of the causes of problems or student success.

In conclusion, assessment in education is a systematic and continuous process aimed at collecting, analysing, and interpreting information about student learning outcomes to improve the quality of education. Beyond merely administering tests, assessment encompasses a broader scope, including identifying learning objectives, monitoring progress, and diagnosing challenges to provide meaningful feedback for both students and educators. Effective assessment aligns with curriculum standards, employs diverse techniques and instruments, and involves stakeholders in creating a comprehensive understanding of student achievements and learning needs. Consequently, assessment serves as a crucial tool for decision-making, fostering positive learning environments, and enhancing the overall teaching and learning process.

#### **a. Assessment of Learning**

Assessment of learning (*AoL*) refers to the measurement of learning using assessments. In addition, Şahin et al. (2023) explain that assessment of learning is an assessment method used to improve the learning process and teaching effectiveness. Meanwhile, Mustapha et al. (2020) defines assessment of learning as a final exam that aims to determine the level of achievement of certain competencies in various specific fields (English, science, mathematics and others). The previous explanation concluded that assessment of learning refers to strategies designed to determine what students know, indicate whether students have achieved the curriculum outcomes or goals of the student's individual program or not, or to indicate proficiency and make decisions about the student's future program or placement. Assessment of learning is an assessment that is common and produces statements or symbols of how well students learn (ÇÜM, 2021). In other words, assessment for learning is a process that summarizes and reports on student achievement at a specific point in time and is also known as "*summative assessment*".

This information provides teachers, school managers, parents and students with a reliable and sound summary of progress and achievement. According to Logue et al. (2025) assessment for learning provides evidence of student learning and this evidence can be used to inform reporting, differentiation and pioneering. Assessment for learning focuses on groups and groups of students, schools and teachers, and stakeholder interests (Daud et al., 2023). Furthermore, the role of assessment for learning as a summative assessment helps summarize what students know, understand or can do against relevant year level achievement standards for different areas or subjects to report on progress. In addition, assessment for learning provides information on school accountability while maintaining the autonomy of schools to continue school-based decision making. Thus, assessment of learning is designed to provide evidence of achievement to parents, other educators of the students themselves and sometimes to outside groups such as employers and other educational institutions.

Assessment of learning emphasizes the purpose of student privilege, structured quantification of student learning and is carried out for students. Furthermore, Hamdu et al. (2020) added that assessment of learning is usually carried out at the end of the learning or syllabus cycle, feedback to students who have a less clear influence on student learning compared to assessment for learning. However, students rely on their grades and teacher comments and to make decisions about their future learning efforts. Thus, the quantification results of learning outcome measurements should be used to provide students with feedback to be used as a consideration in deciding future learning. Assessment of learning aims to measure, validate and report the level of student learning so that logical decisions can be made (Şahin et al., 2023). Assessment of learning provides valuable information that helps teachers make judgments about student achievement, informs educational decisions for parents, students, and institutions, and supports program review and revision by educational leaders.

### **b. Assessment for Learning**

Assessment for learning focuses on classroom strategies and techniques that are relevant to classroom learning. Furthermore, Kivunja (2015) have identified two sequential actions that can make assessment practices “formative” in that students need to become aware of gaps in their current abilities relative to desired goals and need to take action to address those gaps. This explanation outlines formative assessment as: (a) the selection of tasks that are aligned with goals and have the potential to reveal gaps; (b) open-ended student conversations; (c) the use of thinking-provoking questions; (d) the judicious use of tests; (e) the quality of feedback; and (f) engaging students in peer and self-assessment. Thus, as assessment for learning is enacted within the teacher education community and in policymaking in English-speaking countries, it appears that this pedagogical approach to assessment for learning has become widely accepted.

Teachers use their knowledge of the “gaps” to provide timely feedback to students on how they can close the gaps. By embedding assessments that elicit student explanations formally within the unit, as part of the lesson plan, or during “quick” teaching moments, teachers take the opportunity to close gaps in student understanding. Consequently, student learning is expected to increase (Wagner et al., 2024). In other words, the classroom teaching context is the setting in which assessment for learning occurs.

Assessment for learning is part of teaching and learning and is indeed a powerful form of learning. This seems somewhat different from an evaluative approach to assessment where there are no agreed targets or goals (Şahin et al., 2023). This may be an extreme position that will seem foreign to many, but the call for this open approach to assessment where learning is freed from the constraints of standards, outcomes or teachers having a significant influence is perhaps most prominent in educational circles.

Assessment for learning is known to be the only form of formative assessment available for education. There are a variety of themes of using assessment to support learning outcomes rather than simply as summative judgments about students, teachers or schools. According to Brown & Peterson (2009) explain that in New Zealand for example computer-administered tests are used for diagnostic and accountability purposes and do so effectively because they are controlled by the school rather than externally administered and provide teachers with information about “*which learners need and what next*” (Brown & Hattie, 2012). Meanwhile, Crooks (2010) explained that several data collection techniques, direct observation of performance, portfolios, long-constructed response products as other approaches are widely used in New Zealand secondary schools and systematic ways to ensure the validity and reliability of assessments including the use of multiple assessments for student work, external validation of ratings and the use of assessment rubrics with specified assessment criteria.

### **c. Assessment as Learning**

Recognizing the gap between what learners currently know and what they need to know forms the focus of formative assessment. According to Retnawati et al. (2024) states that ensuring that each student understands his or her own learning gaps is essential to learning. Furthermore, Logue et al. (2025) suggests that this will include some use of self-assessment. It is interpreted as encompassing all activities undertaken by teachers and/or their students that provide information to be used as feedback to modify the teaching and learning in which they are engaged.

Assessment for learning in some teaching and learning contexts embodies the spirit of *AaL*, but in others it simply fits the “message”. They add that there is a difference between assessment for learning as a tool to look at past performance and to help prepare for future learning (Luo & Chan, 2022). They acknowledge that in some cases the emphasis on assessment for learning has become so formulaic that it does little to help us look back. Formative assessment processes are acknowledged to have had a potentially “deformative” impact and are reduced to a mechanistic approach to assessment for learning, where assessment becomes highly restricted, limiting the focus on learning.

Concepts such as assessment as learning are being reinterpreted to account for dominant discourses where the agenda of performativity and accountability prevails. The position of assessment as learning within assessment and learning is not clear (Akar & Uluçınar, 2023). Furthermore, assessment as learning is not only understood as an adjunct to teaching and learning but also as a process through which students engage in assessment where this can be presented as part of learning or assessment as

learning. The development of student engagement and responses to student self-assessment with a focus on exploration processes such as self-regulation, self-efficacy, metacognition and feedback as dimensions of assessment and learning are believed to be an important part of the explanation (Danni et al., 2024). However, encourages a re-shaping of the understanding of assessment practices and places assessment as learning as an important foundation for assessment of learning. Thus, it can be concluded that assessment as learning is a complex interaction between assessment, teaching and learning which basically states that students must understand their own learning progress and goals through several processes that are in cognitive events within themselves.

The implicit meaning is the need for students to continue to be active in both learning and assessment. This indicates how students are involved in assessment. In line with Danni (2000) student self-assessment is the heart of assessment as learning. However, there is a special emphasis on social constructivism as a theoretical basis in this case. Assessment as learning emphasizes how learning and learning. Thus, the relationship between teaching, learning and assessment is very difficult to define in the sense that has been explained.

The fact is that social constructivism fails to provide an adequate explanation. Part of the description developed extends theoretical insights into the theory of cultural historical activity. Without a theoretical attitude expressed by clearly, the concepts being explored will fall victim to whatever dominant discourses may invade the terminology. Guida et al. (2018) has argued that despite decades of debate about the precise theoretical underpinnings of formative assessment, and how this in turn may influence its development, it is essentially a behaviourist, goal-based model that is emerging in practice. The realities of classroom practice and the powerful role of mediation will continue to create authentic experiences for assessment and learning.

In conclusion, formative assessment focuses on identifying and addressing the gap between what learners know and what they need to know, emphasizing student understanding of their own learning gaps and the use of self-assessment. While formative assessment aims to provide feedback for modifying teaching and learning processes, its implementation often becomes mechanistic, limiting its potential to enhance learning. The concept of assessment as learning highlights active student engagement through self-regulation, self-efficacy, and metacognition, fostering an understanding of their learning progress and goals. However, the theoretical foundation of formative assessment remains contested, with influences from social constructivism and cultural-historical activity theory shaping its interpretation. Despite these complexities, the interplay between teaching, learning, and assessment underscores the need for a more dynamic and authentic approach to support meaningful educational experiences.

In this context, the researcher will focus on the role of summative assessment, particularly the development of assessment items, to effectively measure student learning and achievement. Summative assessment, also known as assessment of learning (*AoL*), is a critical tool for evaluating students' overall performance at the conclusion of a learning period. It serves as a final judgment of what students know and can do, aligning with specific learning goals or curriculum outcomes. The primary purpose of summative assessments is to provide a clear picture of student progress, helping to inform decisions related to future learning pathways, course placements, or even employment opportunities.

Given the significance of summative assessments in reporting academic achievements, the development of assessment items that are valid, reliable, and aligned with learning objectives is crucial. A well-constructed summative assessment can provide valuable evidence of student proficiency, making it essential for teachers and educational stakeholders to design questions that accurately reflect the learning outcomes. The goal is to ensure that the assessment captures a comprehensive view of the students' skills, including critical thinking, problem-solving, and application of concepts, all of which are necessary for academic success.

Based on the various theoretical perspectives above, assessment in education is understood as a systematic, continuous, and comprehensive process that not only serves to measure learning outcomes but also serves as a tool to improve the quality of learning. Assessment encompasses a series of activities, from collecting, analyzing, and interpreting information related to student learning outcomes to support informed educational decision-making. In this context, assessment cannot be simplified as a

test, but rather is a broader process involving active interaction between teachers and students, including providing feedback, reflection, and the ongoing development of understanding.

Furthermore, assessment is closely linked to learning objectives and curriculum achievement. Effective assessment must be able to reflect the extent to which students have achieved competencies, including knowledge, skills, and attitudes. This demonstrates that assessment has a strategic function not only as an evaluation tool but also as a diagnostic and learning development instrument. Therefore, assessment must be systematically designed using various techniques and instruments that are valid, reliable, and appropriate to the characteristics of learning.

In practice, educational assessment is classified into three main approaches: assessment of learning, assessment for learning, and assessment as learning. Assessment of learning focuses on measuring learning outcomes at the end of the learning process, summative in nature, and is used for reporting and academic decision-making. Meanwhile, assessment for learning emphasizes the formative function of assessment, which aims to provide feedback for continuous improvement of the learning process. Assessment as learning positions students as active subjects engaged in reflection and self-regulation of their learning through assessment activities.

Although these three approaches have complementary roles, in practice, assessment of learning often dominates, focusing on final results, while the formative and reflective functions of assessment are not optimally utilized. This situation tends to make assessments administrative in nature and contribute less significantly to improving the quality of learning. Furthermore, the complexity of the assessment process requires teachers to be competent in designing instruments capable of measuring various aspects of competency holistically, a challenge in many educational institutions.

Based on this synthesis, this study positions assessment as a strategic instrument that functions not only as a tool for measuring learning outcomes but also as a means to develop the quality of learning in a contextual and meaningful manner. However, this research specifically focuses on developing assessments of learning in the form of summative assessments, particularly on developing valid, reliable, and contextualized test items. In addition, summative assessments, which have tended to be limited to measuring cognitive aspects, need to be developed to be more comprehensive by integrating higher-order thinking skills, problem-solving, and application skills in real-world contexts. Therefore, the development of assessment models in this research is aimed at producing instruments that not only accurately measure learning outcomes but also reflect students' competency needs in accordance with curriculum and workplace demands.

Therefore, the quality of learning is highly determined by the quality of the assessments used. Appropriately designed assessments will provide a comprehensive picture of students' abilities and serve as a basis for increasing learning effectiveness. Therefore, the development of innovative, contextual, and needs-based assessment models is imperative in efforts to improve the quality of education, particularly in the context of vocational education.

## **1. Classical Test Theory Concept**

Classical test theory, also known as pure test theory, that systematic differences in test takers' answers are caused only by variations in the abilities being measured. In addition, this theory assumes that external and internal factors of test takers are constant with strict standards or have random effects (Istiyono, 2020). In this theory, the observed score ( $X$ ) is considered as a combination of the pure score ( $T$ ) and the error score ( $E$ ), with both scores being independent of each other (Steyer, 2001).

Classical test theory states that a pure score can be obtained if there is no error in measurement. However, limitations of measuring instruments and measurement approaches often cause the observed score ( $X$ ) to differ from the actual ability ( $\theta$ ), where this error is assumed to be a random variable with a normal distribution (Cappelleri, et.al. 2014). As a result, classical test theory suggests that measurement results are not completely accurate, so that the score obtained by a person may not represent the actual score (Vincent & Shanmugam, 2020). In other words, repeating the same test should not change the actual score. The observed score ( $X$ ) almost always reflects the actual score ( $T$ ) which is affected by measurement error ( $E$ ), where this error can increase or decrease the observed score (Adedoyin, et al., 2008). The magnitude of measurement error is measured through the standard

deviation of the random error distribution of everyone. In classical test theory, it is assumed that the distribution of random errors is uniform for all individuals. Therefore, the standard deviation of the error, known as the standard error, is used as the main indicator in classical test theory (CTT).

Instrument analysis methods based on classical test theory have long been used to evaluate tests. The focus in this analysis includes the total test score, the frequency of correct answers to assess the level of difficulty of the questions (b), response frequency analysis to detect interference, test reliability, and item-total correlation to evaluate item discrimination (a) (Abedalaziz & Leng, 2013; Impara & Plaka, 1997; Hambleton, 2000). One of the significant weaknesses of classical test theory is its dependence on samples, so that statistics describing items and questions are greatly influenced by sample characteristics (Hambleton, 2000). This criticism may be less relevant if the sample used is representative enough and does not change significantly over time. However, this still needs to be verified, and complex strategies must be designed to overcome these limitations.

## 2. Classical Test Theory Analysis

Classical test analysis uses traditional items and sample-dependent statistics. This includes item difficulty and item discrimination estimates, distractor analysis, item correlations and various related statistics (Retnawati, 2016). Most psychometric analyses focus on assessing examinees at the test score level rather than at the item level. Classical test analysis also generally includes measures of score reliability, item difficulty and instrument discrimination. Item analysis is a set of statistical procedures that focus on selecting items to maximize score reliability. The main classical analysis statistics are item difficulty, item discrimination and instrument reliability.

### a. Item Difficulty

Item difficulty in classical test theory (CTT) is one of the characteristics of an item. Item characteristics are the proportion of examinees who take the test and those who answer the item correctly. The greater the percentage of items answered correctly by examinees, the easier the item is. The higher the difficulty value, the easier the item is to understand. The item difficulty index is calculated by dividing the number of examinees who answer the item correctly by the total number of examinees who answer the item. Items answered correctly by 75% of the test takers will get a difficulty index or p value of 0.75. Meanwhile, items answered correctly by 40% of the test takers will get a lower difficulty index or p value of 0.40 (Matlock-Hetzel, 1997). Item difficulty is denoted as p and symbolically given as:

$$P = \frac{R}{N} \quad (8)$$

Where:

P: Difficulty level

R: The number of examinees who answered the questions correctly

N: Total number of examinees

**Table 1.** Difficulty Level

| Difficulty Level (p) | Interpretation |
|----------------------|----------------|
| $P \leq 0,30$        | Difficult      |
| $0,31 \leq 0,70$     | Average        |
| $P > 0,70$           | Easy           |

Tests in general should have an item difficulty level of around 0.5, although individual items can range from 0.2 to 0.8. Similarly, Ashraf and Jaseem (2020) state that items with a difficulty level of around 0.5 are generally good, but there is no hard and fast rule for acceptable variation from this benchmark. For most test items, a difficulty level between 0.3 and 0.7 maximizes information about individual differences among test takers.

### b. Discrimination Index

The standard deviation of observed scores and test reliability are often used to estimate the standard error of measurement. The level of this error affects the accuracy of psychological attribute measurement, especially the actual score (T) (Adedoyin, et al., 2008). In addition, the standard error of

measurement is used to build a confidence interval around a certain score.

Instrument analysis methods based on classical test theory have long been applied in test evaluation. This approach focuses on the total test score, the frequency of correct answers as an indicator of the level of difficulty of the question (*b*), response frequency analysis to identify anomalies, test reliability, and item-total correlation to assess the level of item discrimination (*a*).

However, the main weakness of classical test theory lies in its dependence on the sample. Statistics that describe test items and questions are greatly influenced by the characteristics of the sample used (Hambleton, 2000). However, these weaknesses can be minimized if the samples used are sufficiently representative and consistent over time. However, verification is still needed, and more complex strategies must be developed to overcome these limitations.

### 1) Dichotomous Data of Discrimination Index

The item discrimination index is used as a simple measure to evaluate the ability of an item to differentiate the abilities of test takers, by utilizing extreme group analysis. The calculation of the discrimination index *D* is carried out through the following steps: (*a*) sorting participants based on their test scores, and (*b*) separating the 27% of participants with the highest scores and the 27% of participants with the lowest scores for analysis. (Istiyono, 2020) explained that the value of 27% was chosen because it was able to maximize the differences in the normal distribution while ensuring sufficient data to analyze the discrimination index (*b*), which is further explained as follows:

$$D = P_u - P_l \quad (9)$$

Where:

*D*: Discrimination index item

*P<sub>u</sub>*: Proportion of correct answers for the top group

*P<sub>l</sub>*: Proportion of correct answers for the lower group

The interpretation of the differential power index can be seen in the following table.

**Table 2.** Discrimination Index

| Discrimination Index                  | Items Quality                                      |
|---------------------------------------|----------------------------------------------------|
| $D \geq 0.40$                         | Item is functioning satisfactorily.                |
| $0.30 \leq D \leq 0.39$               | Good item; little or no revision required.         |
| $0.20 \leq D \leq 0.29$ $D \leq 0.19$ | Bad item; should be removed or revised completely. |

Meanwhile, the item discrimination coefficient includes two indicators of the effectiveness of item discrimination which include point-biserial correlation and biserial correlation coefficient. The choice of correlation depends on the type of question to be answered. One of the main disadvantages of the *D* discrimination index is that only the upper 54%-27% and lower 27% are used to calculate item discrimination and 46% of the examinees are ignored. Therefore, the advantage of using the item discrimination coefficient in determining item discrimination over the item discrimination index is that every examinee who takes the exam is used to calculate the item discrimination coefficient.

### 2) Polytomous Data of Discrimination Index

Several methods are used to calculate the discrimination index for polytomous data. The most common approaches include the polyserial correlation coefficient and the generalized discrimination index (GDI). Furthermore, the polyserial correlation coefficient measures the relationship between polytomous items and the overall test score. This is particularly useful when dealing with ordered categories.

$$r_{ps} = \frac{\sum_{k=1}^m (P_k \bar{X}_k - \bar{X})}{S} \sqrt{\frac{n}{\sum_{k=1}^m P_k (1 - P_k)}} \quad (10)$$

Where:

*P<sub>k</sub>*: Proportion of test takers in each category *k*

$\bar{X}_k$ : Average total score of test takers in each score category *k*

$\bar{X}_k$ : Average total score of all test takers

$S$ : Standard deviation of total test scores

$n$ : Total number of test takers

The polyserial correlation coefficient is advantageous because it takes into account the ordinal nature of polytomous items, thus providing a more accurate measure of item discrimination. Meanwhile, the general discrimination index (GDI) compares the average scores of test takers across ability groups for each category.

$$GDI = \frac{\sum_{j=1}^m W_j (\bar{X}_j - \bar{X})}{m} \quad (11)$$

Where:

$W_j$ : Category item  $j$

$\bar{X}_j$ : Average total score of test participants in category  $j$

$\bar{X}$ : Average total score of all test participants

$m$ : Number of categories

This method provides a comprehensive way to assess the discriminatory power of test items by considering the weighted score differences in each response category. For this method to be used optimally, researchers need to first group the responses appropriately and calculate the required statistical parameters. The discrimination index obtained can help identify test items that are able to effectively differentiate the ability levels of participants (Retnawati, 2016). Items with low discrimination indices need to be considered for revision or replacement so that the overall quality of the assessment can be improved.

In the context of polytomous data, discrimination indices become very important tools for test development and evaluation. By applying techniques such as polyserial correlation coefficients and general discrimination indices, researchers can gain better insight into the effectiveness of test items, as well as ensure that the assessments carried out truly reflect the abilities of test participants. This approach significantly improves the accuracy and reliability of measurements, which ultimately supports the creation of more robust and valid tests in the fields of education and psychology.

### c. Pseudo-Guessing

One of the important parameters in classical test theory analysis is the evaluation of distractor function. According to Asamoah and Ocansey (2019) stated that the effectiveness of distractors is measured based on the number of test takers who choose them. Similarly, Musa et al. (2018) identified two main criteria for assessing distractor function, namely ability to choose and discrimination power. First, a distractor is considered functional if it is chosen by at least some test takers, with a minimum recommendation of 5% as the limit for maintaining it. If no participants choose the distractor, then the distractor is considered ineffective and needs to be replaced. Second, a distractor is considered to have good discrimination ability if it is chosen more often by low-ability participants compared to high-ability participants. This shows that effective distractors have a negative correlation with the total test score. Conversely, distractors with a positive or zero correlation to the total test score are considered ineffective and can reduce the overall quality of the test items.

Classical Test Theory (CTT) is one of the fundamental approaches in measuring and evaluating test instruments. CTT focuses on the relationship between observed scores, true scores, and measurement errors, if observed scores consist of pure scores (true scores) plus random measurement errors. This method is used to evaluate the reliability, validity, and function of test items, and to identify the quality of distractors in multiple-choice tests. In this study, CTT analysis will be conducted through a limited trial to obtain information on the quality of the test instrument, including item discrimination, difficulty level, test reliability, and distractor effectiveness. The limited trial was conducted as an initial stage to obtain empirical data used to identify weaknesses in test items, evaluate measurement accuracy, and provide recommendations for instrument revision. This step is expected to ensure that the developed instrument meets the required reliability and validity standards.

## 4. MATERIALS AND METHODOLOGY

This study employed a quantitative evaluative research design aimed at examining the psychometric quality of an existing English Test for Tourism using the Classical Test Theory (CTT) framework. The study focused exclusively on analyzing the empirical quality of the test based on students' response data. The primary objective was to evaluate item characteristics and overall test reliability to determine whether the test meets acceptable standards for educational measurement and assessment. The participants consisted of vocational high school students enrolled in tourism programs in Indonesia who had completed the English Test for Tourism as part of their academic assessment. A total of 90 student response sheets were included in the analysis. The sample size was considered adequate for Classical Test Theory analysis, which generally requires enough respondents to obtain stable estimates of item difficulty, discrimination, and reliability indices.

The instrument analyzed in this study was a multiple-choice English Test for Tourism designed to measure students' English proficiency within tourism-related contexts. Each item consisted of four answer options, with one correct response. The test covered competencies relevant to tourism English, including vocabulary in hospitality contexts, functional expressions used in service interactions, reading comprehension related to tourism scenarios, and grammar used in communicative tourism situations. All items were scored dichotomously, with 1 assigned to correct answers and 0 to incorrect answers. The total test score was obtained by summing the scores across all items.

The data were collected from students' responses during regular classroom assessment sessions conducted under standardized testing conditions. All answer sheets were compiled and coded into a dataset prior to statistical analysis. Incomplete response sheets were excluded to ensure data accuracy and consistency. The dataset was then prepared for item analysis using the Classical Test Theory approach. Data analysis focused on both item-level and test-level statistics. Item difficulty was calculated as the proportion of students who answered each item correctly, with values ranging from 0 to 1. Items were classified as difficult when the difficulty index was below 0.30, moderate when it ranged between 0.30 and 0.70, and easy when it exceeded 0.70. Item discrimination was calculated using the upper and lower group method, in which the top 27% and bottom 27% of students were compared. The discrimination index indicated how effectively each item distinguished between high-performing and low-performing students. Items with discrimination values below 0.20 were considered poor, while higher values indicated better discriminating power.

In addition, distractor analysis was conducted to evaluate the effectiveness of incorrect answer options. A distractor was considered functional if it was selected by at least 5% of students and attracted more responses from lower-performing students than from higher-performing students. Non-functioning distractors were identified for potential revision. Test reliability was estimated using the Kuder-Richardson Formula 20 (KR-20), which is appropriate for dichotomously scored items. A reliability coefficient of 0.70 or higher was considered acceptable for educational assessment purposes. Descriptive statistics, including the mean, standard deviation, minimum score, and maximum score, were calculated to provide an overview of students' overall performance. All data were treated confidentially and analyzed anonymously, and the results were used solely for research purposes to evaluate the psychometric quality of the English Test for Tourism.

## 6. RESULTS

English language proficiency measurement model for vocational high school students in the field of tourism services has been developed to adapt to school learning indicators. The developed English language proficiency measurement model comprises five measurement factors: (a) Meeting New People; (b) Introducing Others; (c) Giving Directions; (d) Descriptive Text; and (e) Narrative Text. These measurement factors are supported by the following conceptual definition of cultural insight. Therefore, cultural insight is understood as a symbol of norms, beliefs, and cultural practices passed down from generation to generation and serving as a guideline in the social life of the community. The following experts provide different emphases on cultural insight, which serve as the source for developing an English language proficiency measurement model based on cultural insight in English.

The English language test instrument based on Cultural Insight in English was tested through analysis

stages. The first stage was a content validity analysis consisting of seven expert judgments in the areas of measurement, grammar, and English. Furthermore, a Unidimensional Confirmatory Factor Analysis was conducted to provide evidence of the instrument's construct validity. The third stage was a Classical Test Theory analysis to assess the quality of the developed items. R 4.4.2 software was used to analyze CFA, reliability, and CTT. This study began with an analysis of content validity. Seven expert judgments tested 30 items for validity. The Aiken-V results are shown as follows:

This transformation has endured and remains preserved to this day, with the figure of San Sebastián serving as a highly symbolic and integral component of the celebration, associated with a specific moment, space, and ritual.

**Table 1:** Validity Category Aiken's V

| Validity Category | Aiken's V | Number of Items | P?    | Decision        |
|-------------------|-----------|-----------------|-------|-----------------|
| High              | $\geq 80$ | 17 Items        | 56.7% | Valid           |
| Moderate          | 0.60-0.79 | 5 Items         | 16.7% | Must be Revised |
| Low               | $< 0.60$  | 8 Items         | 26.6% | Invalid         |
| Total             | -         | 30 Items        | 100%  | -               |

The Cultural Insight in English (CIE) instrument was initially developed with a total of 30 items designed to represent English language competency within the context of cultural insight. All items then underwent content validation by 7 experts, including two experts in English language education, two experts in measurement, and three experts in grammar, to ensure the items' appropriateness for the construct being measured. Content validity was analysed using the Aiken's V index, which assesses the level of inter-expert agreement on the relevance, clarity, and representativeness of each item. Aiken's V was chosen because it is appropriate for polytomous assessment scales and requires more than two raters, thus providing a more stable estimate of content validity. The analysis showed that of the 30 validated items, 17 (56.7%) were in the high validity category, with an Aiken's V value of  $\geq 0.80$ . These items demonstrated a very strong level of expert agreement and were deemed to accurately and clearly represent the construct of Cultural Insight in English. In addition, 5 items (16.7%) fell into the moderate validity category, with Aiken's V values between 0.60 and 0.79. Items in this category were deemed conceptually relevant but required minor improvements, particularly in terms of wording, cultural context, or clarity of the item stimulus. Conversely, 8 items (26.6%) demonstrated low content validity, with Aiken's V values below 0.60. These items reflected low inter-expert agreement, thus being deemed poorly representative of the construct being measured and deemed methodologically unfit for retention.

Based on these results, an item selection process was conducted, retaining all items in the high category and items in the moderate category that had been revised based on expert input. Meanwhile, all items with low content validity were eliminated from the instrument. Through this process, the number of instrument items was reduced from the initial 30 to 22 items that were deemed content valid. In addition, the instrument, with 22 selected items, was deemed to meet the principles of content representativeness and construct suitability, making it suitable for use in the empirical testing phase. Thus, the content validation results indicate that the Cultural Insight in English instrument has a strong conceptual foundation and has undergone a systematic development process. Therefore, the 22 items version of the instrument is ready for use in a limited trial to obtain further empirical evidence of validity and reliability.

The next stage was a CFA analysis using a unidimensional CFA approach to test the measurement model formulated based on theory. The tested model consisted of 22 items that simultaneously loaded onto a single general factor. This approach allows for evaluation of each item's contribution to the general construct. The analysis was conducted using R software version 4.4.4 in the R Studio environment to ensure that the proposed model structure optimally explains the relationships between latent variables in accordance with the established theoretical framework. Table 11 below displays the results of the CFA analysis.

**Table 2:** Kaiser–Meyer–Olkin (KMO) test and Bartlett's Test of Sphericity

| Nu. | Aiken's V                                             | Point   |
|-----|-------------------------------------------------------|---------|
| 1   | Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy | 0.728   |
| 2   | Bartlett's Test of Sphericity (Chi-square)            | 435.566 |

PSYCHOMETRIC ANALYSIS OF A TOURISM  
ENGLISH TEST IN INDONESIA: A CLASSICAL  
TEST THEORY APPROACH

|   |           |         |
|---|-----------|---------|
| 3 | df        | 153     |
| 4 | (p-value) | > 0.001 |

The Kaiser–Meyer–Olkin (KMO) test showed a value of 0.728, which is in the middling category. A KMO value exceeding the minimum threshold of 0.50 indicates that a substantial proportion of inter-item variance can be explained by the latent factor, thus ensuring sufficient sample size for factor analysis and modelling. Therefore, the inter-item correlation in the Cultural Insight in English (CIE) context in English language proficiency assessment instrument meets the statistical requirements for further analysis. Furthermore, the Bartlett’s Test of Sphericity showed a Chi-square value of 435.566 with 153 degrees of freedom (df) and a significance level of  $p < 0.001$ . These results indicate that the inter-item correlation matrix differs significantly from the identity matrix, indicating a strong correlation between the items in the instrument. In other words, the items do not stand alone but are interrelated in measuring the same latent construct cultural insight-based English language proficiency.

Thus, the combination of adequate KMO values and significant Bartlett’s Test results provides empirical evidence that the research data meets the assumptions of sampling adequacy and the appropriateness of the correlation structure. Therefore, the data are suitable for further analysis, both within the framework of construct validity through CFA and measurement modelling using Item Response Theory (IRT) in the final trial stage.

**Table 3:** Standardized Factor Loadings of Reading Indicators

| No. | Item | Factor Loading | Value |
|-----|------|----------------|-------|
| 1   | B1   | 0.048          | 0.675 |
| 2   | B2   | 0.487          | 0.000 |
| 3   | B3   | 0.239          | 0.028 |
| 4   | B4   | 0.408          | 0.000 |
| 5   | B7   | 0.375          | 0.000 |
| 6   | B8   | 0.655          | 0.000 |
| 7   | B9   | 0.170          | 0.129 |
| 8   | B10  | 0.655          | 0.000 |
| 9   | B11  | 0.270          | 0.012 |
| 10  | B17  | 0.420          | 0.000 |
| 11  | B18  | 0.414          | 0.000 |
| 12  | B19  | -0.027         | 0.814 |
| 13  | B20  | 0.649          | 0.000 |
| 14  | B21  | 0.024          | 0.833 |
| 15  | B22  | 0.654          | 0.000 |

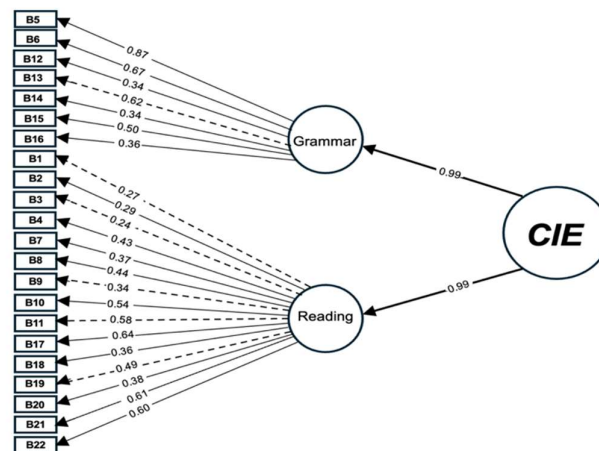
**Table 3:** Standardized Factor Loadings of Reading Indicators

| No. | Item | Factor Loading | Value |
|-----|------|----------------|-------|
| 1   | B5   | 0.604          | 0.000 |
| 2   | B6   | 0.483          | 0.000 |
| 3   | B12  | 0.564          | 0.000 |
| 4   | B13  | -0.082         | 0.474 |
| 5   | B14  | 0.384          | 0.000 |
| 6   | B15  | 0.416          | 0.000 |
| 7   | B16  | 0.348          | 0.001 |

The instrument’s construct validity was tested using Confirmatory Factor Analysis (CFA) to assess the suitability of the theoretical model for the Cultural Insight in English-based English proficiency assessment model using the MCMA format. The initial model tested was a unidimensional, one-factor model. The table above shows eight factors with eigenvalues  $>1$ , with the first factor having a larger eigenvalue (4.71) than the second factor (2.10). These results indicate the dominance of one main factor, making the instrument substantively unidimensional. The presence of several factors indicates noise or interference originating from several items that do not fully represent the main CIE construct. Therefore, the one-factor model remains theoretically relevant but requires refinement at the item level.

Furthermore, the CFA results showed that the factor loadings on the instrument items ranged from - 0.082 to 0.654. Referring to CFA criteria, an item is considered to have good construct validity if the

factor loading is  $\geq 0.40$  and is significant ( $p < 0.05$ ). The table above shows that items with high and significant factor loads ( $\lambda \geq 0.50$ ;  $p < 0.05$ ) include items 5, 8, 10, 12, 20, and 22. These items demonstrate strong representation of the Cultural Insight in English construct and can be categorized as primary indicators or the true face of the construct being measured. Meanwhile, items with moderate but significant factor loads ( $0.30 \leq \lambda < 0.50$ ) are seen in items 2, 4, 6, 7, 14, 15, 16, 17, and 18, which can still be maintained with the note that editorial or contextual improvements are made to the instrument. Furthermore, items with very low factor loads, or negative values, are items 1, 9, 13, 19, and 21, which do not reflect the CIE construct, so they are empirically invalid and need to be revised substantially.



**Figure 1:** CFA result

The results of the second-order Confirmatory Factor Analysis (CFA) indicate that the measurement model of Cultural Insight in English (CIE) ability demonstrates a strong and stable hierarchical structure. The second-order latent construct, namely CIE Ability, is empirically formed by two first-order latent constructs: Reading and Grammar. The very high standardized factor loadings on both paths ( $\lambda = 0.99$ ) indicate that Reading and Grammar are essential and inseparable components in representing the construct of English language ability embedded with cultural insight developed in this study. Furthermore, the strong relationship between CIE Ability and the Reading construct suggests that the ability to comprehend culturally embedded English texts plays a dominant role in shaping students' CIE competence. This finding aligns with the characteristics of English learning in vocational education, particularly in tourism programs, where reading skills are crucial for understanding cultural contexts, tourism-related information, and authentic intercultural interactions. Therefore, Reading can be regarded as a foundational construct in the mastery of CIE. Similarly, the Grammar construct also demonstrates a very strong contribution to CIE Ability. This result confirms that mastery of grammatical structures remains a critical prerequisite for effectively expressing cultural insight through accurate and acceptable English usage. In this context, grammar is not merely understood as the knowledge of formal rules, but as a means of conveying cultural values appropriately and contextually.

At the indicator level, the "Reading" construct is measured by fifteen items that exhibit a range of factor loadings from low to high. Items with high factor loadings, such as B17, B21, and B10, represent more complex reading skills, including the comprehension of implicit meanings and cultural contexts. Meanwhile, items with relatively lower factor loadings were retained due to their theoretical relevance and support from Item Response Theory (IRT) analysis, which indicated that the response categories functioned optimally. The variation in factor loadings among the Reading indicators suggests that this construct captures a broad spectrum of students' abilities, ranging from literal comprehension to interpretative understanding. This strengthens the argument that the CIE instrument is neither narrow nor biased toward a specific ability level, but rather sensitive to differences in students' competencies across multiple proficiency levels. Consequently, Reading functions as an informative and adaptive construct within the measurement of culturally embedded English ability.

The "Grammar" construct is measured by seven indicators that generally show moderate to high factor loadings. Items B5 and B6 demonstrate the strongest loadings, indicating that these items serve as key indicators in representing students' grammatical competence. The strong contribution of these

PSYCHOMETRIC ANALYSIS OF A TOURISM  
ENGLISH TEST IN INDONESIA: A CLASSICAL  
TEST THEORY APPROACH

indicators suggests that the Grammar construct in this model is relatively homogeneous and consistently measures the same underlying latent trait. The presence of several Grammar indicators with moderate factor loadings does not indicate weakness in the model; instead, it reflects the complexity of grammatical mastery within culturally contextualized language use. In English language learning practice, grammatical competence is often acquired progressively and contextually, and therefore variation in indicator contributions enriches the measurement coverage of the construct.

In addition, the results of the second-order CFA provide strong evidence for the construct validity of the CIE instrument. The hierarchical structure confirms that culturally embedded English ability is a multidimensional construct that is systematically organized. The consistent relationships between the second order and first-order constructs indicate that the proposed theoretical model is well supported by empirical data.

In conclusion, the developed CIE instrument meets the criteria for both structural and functional validity. The second-order CFA model not only confirms the appropriateness of the theoretical structure of the instrument but also demonstrates that the instrument is suitable for measuring culturally embedded English language ability among vocational high school students majoring in tourism in an accurate, reliable, and contextual manner.

**Table 4: Reliability Result**

| <b>Dimension</b> | <b>Alpha</b> | <b>Ordinal Alpha</b> | <b>Omega (<math>\omega</math>)</b> | <b>Omega total</b> | <b>Omega Hierarchical</b> | <b>AVE</b> |
|------------------|--------------|----------------------|------------------------------------|--------------------|---------------------------|------------|
| Grammar          | 0.555        | 0.656                | 0.614                              | 0.614              | 0.634                     | 0.305      |
| Reading          | 0.739        | 0.806                | 0.766                              | 0.766              | 0.788                     | 0.281      |

Construct reliability analysis was conducted to evaluate the internal consistency and quality of representation of the latent constructs in the multidimensional CFA model. Reliability was estimated using coefficient alpha, ordinal alpha, omega, and Average Variance Extracted (AVE). The table 15 above shows that the Grammar dimension exhibited a Cronbach's alpha value of 0.555 and an ordinal alpha of 0.656. The ordinal alpha value is more relevant in the context of ordinal data, so 0.656 is the primary reference for interpretation.

Furthermore, the omega value for Grammar, 0.614, indicates a moderate level of reliability. According to Hair et al. (2019), reliability values between 0.60 and 0.70 are acceptable in exploratory research or early-stage instrument development. The hierarchical omega (0.634) indicates that most of the variance in Grammar scores is explained by the hypothesized latent construct, although it does not yet reach the high category. The Grammar AVE value of 0.305 indicates that the construct only explains approximately 30.5% of the variance in its indicators. Conventionally, an AVE value  $\geq 0.50$  is recommended to indicate strong convergent validity (Fornell & Larcker, 1981). However, Fornell and Larcker (1981) also emphasized that an AVE below 0.50 is acceptable if the composite reliability exceeds 0.60, as the construct's reliability is still considered adequate under these conditions. Therefore, even if the Grammar AVE is low, an omega value above 0.60 still allows the construct to be theoretically tenable.

Furthermore, the Reading dimension showed stronger results, with a Cronbach's alpha of 0.739 and an ordinal alpha of 0.806, indicating good internal consistency. The omega value of 0.766 indicates good construct reliability and is above the 0.70 threshold recommended for confirmatory research. The hierarchical omega of 0.788 further confirms that the variance in Reading scores is substantially explained by the latent construct. However, the AVE for Reading of 0.281 is also below the ideal limit of 0.50. This indicates that although internal consistency is good, the proportion of indicator variance explained by the construct is still relatively limited. The low AVE values for both dimensions are likely due to several items with factor loadings below 0.40, resulting in a small contribution of variance to the latent construct. In the context of this study, reliability in the range of 0.60–0.70 is often considered sufficient, especially for new instruments still in the initial validation stage (Nunnally & Bernstein, 1994). In this model, the second-order structure shows a loading of 0.99 from CIE to Grammar and Reading, indicating that both dimensions strongly represent the general construct of CIE. The high second-order loadings indicate conceptual coherence between the dimensions, so that although Grammar has moderate reliability, it is still strongly integrated structurally within the CIE construct. In other words, the reliability results show that the Reading dimension has good psychometric quality,

while the Grammar dimension requires minor improvements in several indicators to improve internal consistency and convergent validity.

The next stage is testing the quality of these items using the Classical Test Theory approach with two parameters: item difficulty and item discrimination. The results of the item quality test are shown in the following table.

**Table 5:** Classical Test Theory Result

| No | Item | Mean  | Difficulty Level | <i>pBis</i> | Discrimination Index | Category        |
|----|------|-------|------------------|-------------|----------------------|-----------------|
| 1  | B1   | 1.427 | Moderate         | 0.062       | Very Low             | Must be revised |
| 2  | B2   | 2.112 | Moderate         | 0.415       | Good Fit             | Good fit        |
| 3  | B3   | 1.281 | Moderate         | 0.238       | Adequate Fit         | Good fit        |
| 4  | B4   | 1.685 | Moderate         | 0.327       | Good Fit             | Good fit        |
| 5  | B5   | 1.787 | Moderate         | 0.495       | Good Fit             | Good fit        |
| 6  | B6   | 1.449 | Moderate         | 0.394       | Good Fit             | Good fit        |
| 7  | B7   | 1.820 | Moderate         | 0.334       | Good Fit             | Good fit        |
| 8  | B8   | 1.753 | Moderate         | 0.594       | Excellent Fit        | Good fit        |
| 9  | B9   | 1.056 | Difficult        | 0.141       | Poor Fit             | Good fit        |
| 10 | B10  | 2.371 | Easy             | 0.487       | Good Fit             | Good fit        |
| 11 | B11  | 1.640 | Moderate         | 0.227       | Adequate Fit         | Good fit        |
| 12 | B12  | 2.247 | Easy             | 0.493       | Good Fit             | Good fit        |
| 13 | B13  | 1.034 | Difficult        | -0.070      | Negative             | Must be revise  |
| 14 | B14  | 1.697 | Moderate         | 0.352       | Good Fit             | Good fit        |
| 15 | B15  | 1.843 | Moderate         | 0.398       | Good Fit             | Good fit        |
| 16 | B16  | 2.022 | Moderate         | 0.360       | Good Fit             | Good fit        |
| 17 | B17  | 2.236 | Moderate         | 0.423       | Good Fit             | Good fit        |
| 18 | B18  | 1.876 | Moderate         | 0.417       | Good Fit             | Good fit        |
| 19 | B19  | 1.000 | Difficult        | -0.020      | Negative             | Must be revised |
| 20 | B20  | 1.764 | Moderate         | 0.564       | Excellent Fit        | Good fit        |
| 21 | B21  | 1.101 | Difficult        | 0.002       | Poor Fit             | Must be revise  |
| 22 | B22  | 1.663 | Moderate         | 0.555       | Excellent Fit        | Good fit        |

Based on the table 16 above, item characteristic analysis based on the Classical Test Theory (CTT) approach was conducted as a preliminary step to evaluate the empirical quality of each item in the Cultural Insight in English (CIE) instrument before conducting further analysis using Confirmatory Factor Analysis (CFA) and Item Response Theory (IRT). The CTT approach was used to identify the mean score, difficulty level, and discrimination index.

Furthermore, based on the analysis of the 22 items tested, the mean score for each item generally ranged from 0.91 to 1.74, with a maximum score of 3. This mean score was then converted into a difficulty index (*p*) by dividing the mean by the maximum score. The conversion results showed that all items fell into the moderate difficulty category ( $0.30 \leq p \leq 0.70$ ). This finding indicates that the CIE instrument has a balanced level of difficulty and is not dominated by items that are too easy or too difficult, thus theoretically being able to optimally measure variations in student ability. The dominance of moderate difficulty levels is an important characteristic of instruments designed for ability measurement, particularly in the context of educational assessment. Items with moderate difficulty tend to provide maximum information about test takers' abilities because they are able to differentiate between low, medium, and high-ability students more proportionally. Therefore, from a difficulty perspective, the CIE instrument meets one of the main criteria for a quality test instrument.

In addition, in terms of discriminatory power, most items demonstrated corrected item-total correlations in the good to excellent category. Several items, such as B2, B5, B8, B10, B12, B15, B16, B17,

PSYCHOMETRIC ANALYSIS OF A TOURISM  
ENGLISH TEST IN INDONESIA: A CLASSICAL  
TEST THEORY APPROACH

B18, B20, and B22, had discriminatory power indices of  $\geq 0.40$ , which is categorized as very good. This indicates that these items have strong ability to differentiate students with high and low abilities on the Cultural Insight in English construct being measured. High discriminatory power confirms that students' responses to the items are consistent with the total test score.

However, the CTT analysis also identified several items with low or even negative discriminatory power, namely items B13, B19, and B21. Negative discriminatory power values indicate that these items are not performing as intended, as students with low ability tend to score higher than students with high ability. Psychometrically, this condition strongly indicates that these items are problematic, both in terms of item construction, stimulus clarity, and appropriateness of scoring. Items B13 and B19, in particular, demonstrated fairly consistent negative discriminatory power values, despite their moderate difficulty levels. This finding strengthens the argument that an ideal level of difficulty does not always guarantee item quality if it is not accompanied by adequate discriminatory power. Therefore, these items are deemed empirically unfit within the CTT framework and require further treatment, either through substantive revisions, scoring modifications, or removal from the pilot instrument.

Item B21 demonstrated very low discriminatory power despite its positive value. This condition indicates that the item's contribution to the overall measurement of the construct is relatively weak. Overall, the CTT analysis results indicate that the majority of items in the CIE instrument meet adequate empirical criteria, both in terms of difficulty and discriminatory power. Of the 22 items analyzed, the majority were categorized as fit and worthy of proceeding to the next testing stage. Meanwhile, a small number of items that did not meet the discriminatory power criteria were clearly identified as requiring revision. Based on the analysis of item characteristics using the Classical Test Theory (CTT) approach with a scoring scale of 0–2, the majority of instrument items generally demonstrated moderate to easy difficulty levels and had adequate discriminatory power. This indicates that most items were able to proportionally represent the variation in student abilities. However, the analysis also revealed that four items needed revision, particularly due to their low or negative discrimination index values and their weak contribution to the instrument's overall reliability. These items, categorized as "unfit," indicated that student responses to the items were not aligned with the latent ability construct being measured.

Therefore, the four items identified as unfit were not immediately eliminated but were instead substantially revised based on expert input, indicator analysis, and cultural contextual suitability within the Cultural Insight in English framework before being tested on a large scale. By making improvements at this stage, the instrument is expected to demonstrate increased discriminatory power, stable reliability, and stronger construct fit in subsequent testing. Thus, the revision of the four items that did not fit became an integral part of the effort to produce a valid, reliable, and suitable cultural insight-based English language ability assessment instrument for use on a wider scale.

The quality of the English ability assessment model within the Cultural Insight in English (CIE) assessment model was demonstrated through a series of psychometric tests reflecting modern assessment principles. High content validity, based on the *Aiken-V* index, indicates expert agreement on the relevance of the items to the construct being measured. According to (Istiyono (2020) Kardanova, et al. (2024), content validity is a crucial step in ensuring that an instrument accurately represents the intended competency domain. In the context of English ability assessment model within the Cultural Insight in English (CIE), content validity becomes even more important because potential cultural bias can influence item interpretation.

Furthermore, confirmatory factor analysis results showed a unidimensional structure, support the assumption that English ability assessment model within the Cultural Insight in English (CIE) constitute a unified construct. This finding aligns with Thi & Trang (2023), who assert that intercultural communicative competence should be understood as a holistic ability, not a collection of separate sub-skills. Thus, the factor structure of this instrument strengthens the theoretical foundation that language and culture are integral to communication practice.

High reliability values (Cronbach's Alpha  $> 0.7$ ) indicate good internal consistency of the instrument. In addition, adequate reliability indicates that items consistently measure the same construct (Natanael & Ramdani, 2021; Retnawati, Apino, et al., 2024; Şahin et al., 2023). This finding supports the research of Novita et al. (2022), which stated that language assessments in vocational schools often

experience reliability issues due to a lack of contextual appropriateness. In contrast, the English ability assessment model within the Cultural Insight in English (CIE) especially instrument demonstrated measurement stability despite integrating complex cultural contexts. The analysis of item quality based on the CTT, which showed a predominance of moderate difficulty and good discrimination, supports Danni et al., (2024), Şahin et al. (2023) view that items with these characteristics are most effective in differentiating students' abilities. In other words, this demonstrates that the CIE assessment model is not only theoretically valid but also functions optimally as a measurement tool.

## 7. CONCLUSIONS

The results of the study indicate that the developed model has excellent psychometric and theoretical quality. Content validity tested through expert judgment using the Aiken-V index showed a high level of expert agreement, indicating that the test items accurately represented culturally context-based English language ability indicators. The instrument's construct structure was also proven strong through the results of Confirmatory Factor Analysis (CFA) and Principal Component Analysis (PCA), which showed the dominance of one main factor as a latent construct of CIE-based English language ability. In addition, the instrument's reliability, as indicated by a high Cronbach's Alpha value, indicates good internal consistency. Overall, these findings confirm that the CIE assessment model meets the criteria of being valid, reliable, and suitable for use in the context of vocational education.

**Acknowledgement:** The author gratefully acknowledges the Indonesian Education Scholarship (BPI), the Centre for Higher Education Funding and Assessment (PPAPT), and the Indonesian Endowment Fund for Education (LPDP) for their financial support of this research. The author also extend appreciation to all institutions and individuals who contributed to the completion of this study, either directly or indirectly.

**Conflicts of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Funding:** This research was supported by the Indonesian Education Scholarship (BPI), the Centre for Higher Education Funding and Assessment (PPAPT), and the Indonesian Endowment Fund for Education (LPDP).

**Ethical Approval:** This study was reviewed and approved by the Academic Committee of Universitas Negeri Yogyakarta.

**Informed Consent:** Written informed consent was obtained from all participants prior to their involvement in this study.

**Data Availability:** The datasets generated and analyzed in this study are not publicly available due to confidentiality and ethical considerations related to participant data and research instruments. However, the data may be made available by the corresponding author upon reasonable request for academic purposes.

## REFERENCES

- Danni, R., Retnawati, H., & Yogyakarta, U. N. (2024). *Investigating Test Equating Methods : English Examination Test*. 7(3), 246–257.
- Dieu, T. T., & Dan, T. C. (2023). *European Journal of Alternative Education Studies EFL Students ' Perceptions and Practices on Learning Strategies Towards ESP For Tourism at School of Social Sciences And Humanities ,.* 92–132. <https://doi.org/10.46827/ejae.v8i2.4824>
- Evenddy, S. S., Nurlily, L., Pendidikan, J., Inggris, B., Sultan, U., & Tirtayasa, A. (2022). *Virtual Tour : A Media To Teach*. 14(2). <https://doi.org/10.26418/jvip.v14i2.54744>
- Garing, J., Firdaus, W., Herianah, H., Ridwan, M., Erniati, E., Budiono, S., & Pariela, T. D. (2023). Identifying and Resolving Conflicts Using Local Wisdom: A Qualitative Study. *Journal of Intercultural Communication*, 23(4), 69–81. <https://doi.org/10.36923/jicc.v23i3.156>
- Handayani, M. N., Ali, M., Wahyudin, D., Mukhidin, Pavlova, A., Pavlova, M., The, F. O. R., Brown, M.,

PSYCHOMETRIC ANALYSIS OF A TOURISM  
ENGLISH TEST IN INDONESIA: A CLASSICAL  
TEST THEORY APPROACH

- Heong, Y. M., Sern, L. C., Kiong, T. T., Binti Mohamad, M. M., Alwi, A., Kamis, A., Ismail, B. L. H., Setiawan, A., Hamid, Z., Yusri, K., Azlan, A. L., ... Jenkin, N. (2018). 濟無No Title No Title No Title. *International Journal of Academic Research in Business and Social Sciences*, 11(1), 1–6. <https://doi.org/10.1007/s11159-018-9718-x>  
x%0Ahttp://www.tvetonline.asia/issue6/zolkifli\_etal\_tvet6.pdf%0Ahttp://hdl.handle.net/10072/52268h  
ttp://www.griffith.ed
- Hulu, N. F., Zebua, E. P., Harefa, A. T., & Maru'ao, N. (2024). Developing English Teaching Materials Based on Local Wisdom for the Tenth Grade Students at SMA Negeri 1 Alasa. *Academy of Education Journal*, 15(2), 1459–1478. <https://doi.org/10.47200/aoej.v15i2.2507>
- Islam, U., Raden, N., & Palembang, F. (2025). *The English Assessment At Common Asean Tourism Curriculum (Catc) A. Waris*. 1, 13–17.
- Istiyono, E. (2020). *Pengembangan Instrumen Penilaian dan Analisis Hasil Belajar Fisika dengan Teori Tes Klasik dan Modern* (Edisi Kedu). UNY Press.
- Kotollaku, M. (2024). *English Language , An Interactive Communication Tool Among Tourism Sector Employees and Foreign Tourists*. 5(3), 79–89.
- Matondang, K., Mira, R., Saragih, B., Simamora, Y., Putri, J. H., Arief, K., Matematika, P., Al, U., & Medan, W. (2022). *Pelatihan asesmen kompetensi minimum tingkat smp*. 1(2), 32–35.
- Natanael, Y., & Ramdani, Z. (2021). Pengembangan Instrumen Dan Pengujian Properti Psikometri Skala Moderasi Beragama Di Perguruan Tinggi. *Tatar Pasundan: Jurnal Diklat Keagamaan*, 15(2), 196–208. <https://doi.org/10.38075/tp.v15i2.227>
- Nggawu, L. O., & Thao, N. T. P. (2023). The impact of Communicative Language Teaching (CLT) Approach on Students' Speaking Ability in a Public Indonesian University: Comparison between Introverts and Extrovert Groups. *International Journal of Language Education*, 7(3), 393–413. <https://doi.org/10.26858/ijole.v7i3.50617>
- Novita, D., Permatasari, C., & Safitri, L. M. (2022). *The Development of Interactive English Multimedia in Tourism Vocational Education as an Adaptive Learning Strategy in the Midst of Covid-19 Pandemic*. 1(2), 96–106. <https://doi.org/10.52352/lacultour.v1i2.903>
- Pambayun, N. A. Y., Munadi, S., Arifin, Z., Setiawan, C., & Retnawati, H. (2023). Industrial work culture education in Indonesian vocational high schools: Teachers' perceptions and practices. *Issues in Educational Research*, 33(2), 713–732.
- Pambayun, N. A. Y., Sofyan, H., & Haryana, K. (2020). Vocational high school infrastructure conditions and the challenges in facing the era of literation and industrial revolution 4.0. *Journal of Physics: Conference Series*, 1700(1), 0–8. <https://doi.org/10.1088/1742-6596/1700/1/012068>
- Pelzer, E. (2019). The Potential of Conjoint Analysis for Communication Research. *Communication Research Reports*, 36(2), 136–147. <https://doi.org/10.1080/08824096.2018.1559138>
- Purwati, D., Faruq Ubaidillah, M., & Restall, G. C. (2023). “Sorry, I Can’t Speak”: English Teachers' Challenges of Teaching EFL Speaking in an Indonesian Vocational High School Sector. *Mextesol Journal*, 47(1), 0–2. <https://doi.org/10.61871/mj.v47n1-1>
- Retnawati, H., Apino, E., & Rafi, I. (2024). *From Investigating the Alignment of A Priori Item Characteristics Based on the CTT and Four-Parameter Logistic ( 4-PL ) IRT Models to Further Exploring the Comparability of the Two Models*. November. <https://doi.org/10.7275/pare.2043>
- Retnawati, H., Kardanova, E., Sumaryanto, Prasajo, L. D., Jailani, Arliani, E., Hidayati, K., Susanti, M., Lestari, H. P., Apino, E., Rafi, I., Rosyada, M. N., Tuanaya, R., Dewanti, S. R., Sotlikova, R., & Kassymova, G. K. (2024). A Systematic Review of the Use of Technology in Educational Assessment Practices: Lesson Learned and Direction for Future Studies. *International Journal of Robotics and Control Systems*, 4(4), 1656–1693. <https://doi.org/10.31763/ijrcs.v4i3.1572>
- Şahin, M. G., Yıldırım, Y., & Öztürk, N. B. (2023). Examining the Achievement Test Development Process in the Educational Studies. *Participatory Educational Research*, 10(1), 251–274.

<https://doi.org/10.17275/per.23.14.10.1>

Suaco, T. P., Mangaliag, A. D., & Gadgad, M. M. (2023). Collaborative Summative Assessment: Means for Enduring Learning and Attainment of 21st Century Skills in the Online Platform. *Journal of Education and Learning*, 12(1), 118. <https://doi.org/10.5539/jel.v12n1p118>

Tambunan, R. W., & Kurniawan, H. (2024). An implementation of integrated performance assessment model in vocational English for communication learning. *JOALL (Journal of Applied Linguistics and Literature)*, 9(1), 167–181. <https://doi.org/10.33369/joall.v9i1.30876>

Thi, T., & Trang, T. (2023). *Needs Analysis About Intercultural Communicative Competence Among Undergraduate Tourism Students*. 2599–2620.

Tsai, F.-H. (2024). Development and Evaluation of an Internet of Things Project for Preservice Elementary School Teachers. *Sustainability*, 16(17), 7632. <https://doi.org/10.3390/su16177632>

Uktolseja, L. J., & Gaspersz, S. (2019). Implementation of chain story techniques to improve english speaking ability at immanuel YPK vocational high school, Sorong City. *Journal of Education and Learning (EduLearn)*, 13(4), 490–497. <https://doi.org/10.11591/edulearn.v13i4.13327>