TECHNIUM
SOCIAL SCIENCES JOURNAL

www.techniumscience.com

Vol. 74/2025

A New Decade for Social Changes

PLUS COMMUNICATION
International Communication & PR

Technium Social Sciences Journal
Vol. 74, 96-112, August, 2025
ISSN: 2668-7798
www.techniumscience.com

Validity and Reliability of Summative Test Assessment for Bachelor of Technology and Livelihood Education Students

Mary Jane A. Moralia[1], Sarah O. Namoco[2]

[1]Faculty, University of Science and Technology of Southern Philippines-Panaon campus, [2]Faculty, University of Science and Technology of Southern Philippines-Cagayan de Oro Campus

maryjane.moralia@ustp.edu.ph, sarah.namoco@ustp.edu.ph

Abstract. Assessment is at the core of the learning process as it enables educators to determine whether learning has truly occurred. One of the essential types of evaluation is the summative assessment, which measures learners' achievements at the end of a course and evaluates their readiness for future academic tasks or professional responsibilities. To ensure fairness and accuracy, summative tests must be both valid and reliable. This study aimed to develop a valid and reliable teacher-made summative test in Curriculum Development and Evaluation, with an Emphasis on the TM2 course, which was administered to 41 BTLEd third-year students. The quality of the test was examined using content validity, reliability, and item analysis, including difficulty and discrimination indices. A Table of Specifications (TOS) was carefully developed based on the course syllabus to guide test item construction, aligning test items with the course's intended learning outcomes, topics, instructional hours, and cognitive levels outlined by Bloom's Taxonomy. Content validity was confirmed through rigorous review by two course experts. Reliability was assessed using the Kuder-Richardson Formula 20 (KR-20), which yielded a coefficient of 0.94, indicating very high internal consistency. Item analysis revealed that most test items had moderate Difficulty and strong discrimination indices. From the initial 160 items, 38 were discarded, resulting in a final 80-item instrument that met established psychometric standards. This study offers a structured process for ensuring the validity and reliability of summative tests, thereby supporting high-quality assessments in teacher education.

Keywords. Content Validity, Item Analysis, Reliability, Summative Assessment, Teacher Education

A. Introduction

Assessment is at the core of the learning process, as it enables us to determine whether learning has truly occurred. It plays an essential role in education by not only evaluating student progress but also guiding curriculum design and teaching approaches [1],[2]. One of the important types of assessment is the summative assessment, which is essential for verifying what learners have achieved at the end of a course and evaluating their readiness for future academic tasks or professional responsibilities According to Dogara et al. [3] and Siti et al. [4], students perceive assessments as a means to secure employment opportunities and demonstrate their acquired skills. The assessment process plays a vital role in teacher education programs,

as it determines how effectively pre-service teachers acquire fundamental pedagogical concepts. Meaningful assessments, along with accurate results, serve as essential components for developing competent educators who create effective curricula through design evaluation and implementation.

Students can demonstrate their learning through three assessment approaches: formative assessment, metacognitive assessment, and summative assessment. Formative assessment provides continuous feedback to students, which improves their understanding while guiding instruction during the learning process [5]. Formative assessment practices incorporate metacognitive assessment, which helps students develop awareness of their thinking processes while also learning effective learning strategies to promote independent lifelong learning [6]. Summative assessment occurs at the end of instructional units to evaluate comprehensive learning outcomes and assess achievement levels [7]. The influence of summative tests on student grades and academic decisions, as well as institutional accountability, remains the highest among the three types of assessments. The assessment method determines all final grades while controlling course completion rates, academic award distribution, and progression to upper levels [7]. The academic progress of students in the Bachelor of Technology and Livelihood Education program depends heavily on summative tests. A well-designed summative assessment serves dual purposes of measuring student learning and providing information about instructional quality. High-stakes assessment instruments used in educational programs for future educators must demonstrate both accuracy and reliability due to their critical nature [8]. Test instruments that are poorly constructed can have adverse effects on student grades and self-confidence, making it essential to validate and ensure reliability in every assessment tool [9].

In the Curriculum Development and Evaluation with Emphasis on TM2 course, these summative tests serve as a final measure of students' understanding of curriculum principles, evaluation strategies, and assessment design, in accordance with the competencies outlined in CHED Memorandum Order No. 75, Series of 2017.This underscores their importance in supporting and assessing the learning process. In line with this, valid and reliable summative tests are needed for ensuring fair and accurate evaluation.

The validity of summative tests determines whether the assessment accurately measures its intended targets, which are the course learning outcomes. The test content demonstrates alignment with its learning objectives through this assessment method [10]. On the other hand, the assessment results demonstrate reliability through consistent outcomes, which occur when different occasions meet or raters evaluate or test administrations take place [1].

The absence of validity or reliability in tests, according to French et al. [10] and Banerjee et al. [11], can lead to unfair grading and decreased student motivation, ultimately resulting in long-term academic and professional setbacks. Newstead [12] cautions that faulty assessment methods cause students to focus on grades rather than actual learning. An assessment requires both validity and reliability to achieve accuracy and value, as noted by [9]. The curriculum-focused courses require students to demonstrate mastery in instructional design, assessment alignment, and curriculum evaluation. Student performance interpretation becomes unreliable when summative tools lack reliability, which may result in unjust academic decisions and missed opportunities for licensure exams or employment.

The Philippine education system faces growing concerns because teacher education programs heavily rely on summative assessments, which lack validity and reliability, and display poor construction [13],[14]. The evaluation process relies on assessment tools to

Technium Social Sciences Journal
Vol. 74, 96-112, August, 2025
ISSN: 2668-7798
www.techniumscience.com

determine grades and guide instruction and academic progression, thus requiring both validity and reliability for a fair and accurate assessment of learning [15]. The absence of essential qualities in assessment tools creates false measures of student competency, which damages the educational objectives [9]. Research by Esmael and Rabut [13], together with Napanoy and Peckley [16], showed that Philippine university tests developed by teachers frequently lack validation procedures and item analysis, which causes assessment tools to deviate from their intended learning outcomes. When performance evaluations fail to reflect actual student competencies accurately, it produces incorrect assessment results, which negatively impact the quality of teaching and learning. According to Austria-Cruz [17], there exists a direct relationship between flawed assessment practices and rising depression rates among Filipino college students. The development of summative assessment tools for professional education courses remains an unaddressed research need, particularly in the Bachelor of Technology and Livelihood Education (BTLEd) specialized program. The assessment tool literature concentrated on general education subjects like English, mathematics, and science [18], [19] without recognizing the distinctive curricular requirements and practical teaching skills in pedagogical courses. Students must achieve high scores in summative evaluations, which assess all learning areas, to succeed in these tests. The lack of suitable assessment tools creates the potential for BTLEd students to receive unfair assessments, which would threaten their readiness to become teachers since their actual competence would remain undetected. The assessment practices in higher learning institutions face critical quality issues because they create a chain reaction that affects teacher education programs.

In response to these concerns, this study aims to develop a valid and reliable teacher-made summative test in the *Curriculum Development and Evaluation, with an Emphasis on the TM2 course.* In addition, the test will be cross-matched with the course syllabus and Table of Specifications, and its validity, reliability, item analysis difficulty index, and discrimination index will be tested. The study aims to advocate for equitable and valid instructionally aligned assessment practices. Furthermore, it addresses the research gap by contributing to the improvement of assessment literacy and academic equity in teacher education. Ultimately, this research supports the goal of preparing future educators who are not only academically competent but also equipped to design effective curriculum and assessment tools in their teaching practice.

Review of Related Literature

Summative Assessment in Teacher Education

Teacher education programs require assessment as an essential component to evaluate learning while also utilizing it for instructional guidance and curriculum development. Higher education institutions use summative assessments to verify student achievement while determining readiness for professional practice [20], [21]. Teacher education programs utilize these assessments to determine if pre-service teachers understand fundamental pedagogical concepts. The implementation of summative tests in educational settings is widespread; however, research shows that these assessments frequently lack both validity and reliability [22]. The studies [7] and [9] demonstrate that summative assessments become ineffective in showing student competencies when they are not correctly aligned with learning objectives and curriculum standards. The urgent need for validated assessments necessitates immediate attention, which is precisely what this study aims to address.

The Southeast Asian and Philippine region faces an increasing problem regarding the standard of summative tests created by teachers. The research conducted by Setiabudi et al. [22]

Technium Social Sciences Journal
Vol. 74, 96-112, August, 2025
ISSN: 2668-7798
www.techniumscience.com

reveal that numerous assessments fail to align with the taught curriculum, raising doubts about their content validity. The existing research on assessment practices focuses primarily on English, Math, and Science subjects [18], [19]. Still, there is limited investigation regarding the validation of summative tests for curriculum-focused courses such as Curriculum Development and Evaluation in the BTLEd program. The existing assessment gap underscores the need for validated assessments that align with instruction to ensure fair evaluation and support the readiness of future educators. The development of a teacher-made summative test for the TM2 course in this study provides a validated assessment model that promotes reliable, curriculum-based assessment practices and will have a substantial impact on future educator preparation.

Importance of Validity in Summative Test

In educational assessment, summative tests serve as critical instruments for determining student learning outcomes, informing educational promotion decisions, and evaluating teaching methods [20]. The definition of validity refers to how well a test measures its designated targets (Messick, 1995). These tests derive their essential value from their ability to deliver fair results and valuable outcomes. According to Kibble [2], validity originates from the interpretations of scores derived from tests rather than from the test itself. Evidence from content, structure, and use consequences establishes the validity of these scores. A valid summative test requires proper alignment with learning objectives as well as accurate curriculum representation to support educational decision-making.

The development and evaluation of summative assessments follow four main types of validity, which include content validity, criterion validity, construct validity, and face validity. According to Hilaldy [23], content validity refers to the process of matching test items to instructional content and objectives. In contrast, criterion validity evaluates how well a test predicts or reflects actual performance. Face validity provides superficial validation, yet it builds stakeholder trust, while construct validity measures the precise assessment of psychological and cognitive traits (Parrott, 1991).

This study employed content validity. The assessment of summative testing heavily relies on content validity because it ensures that test items properly reflect the learning targets defined in the curriculum. Multiple educational-level studies indicate that an inaccurate alignment between test materials and learning objectives produces incorrect student evaluation results. Puspitasari [24] discovered that senior high school summative tests frequently failed to match syllabus content, yet Hilaldy [23] demonstrated that English tests covered only 20% of the material, which created doubts about fairness and instructional validity. The solution to these problems requires experts to establish structured validation methods, which include expert reviews combined with test blueprints and item analysis [25]. The establishment of content validity remains essential for creating summative assessments that provide fair and accurate measures, aligning with instructional objectives.

Importance of Reliability in Summative Test

Reliability is defined as the consistency and precision of a measurement tool or methodology when applied repeatedly under similar conditions. In the study conducted by Ahmed and Ishtiaq [26], it emphasizes that a reliable assessment yields stable and reproducible results over time, across different raters, and within the parts of the instrument itself. For example, a reliable questionnaire should produce similar scores when administered to the same individuals at different times (test-retest reliability) [9], [13], yield consistent results among

Technium Social Sciences Journal
Vol. 74, 96-112, August, 2025
ISSN: 2668-7798
www.techniumscience.com

different evaluators (inter-rater reliability), and have strong internal agreement across its items (internal consistency) [27].

Furthermore, Ahmed and Ishtiaq [26] further highlight that reliability is foundational but not sufficient on its own to guarantee validity. A test can be highly reliable, consistently producing the same result but still invalid if it does not accurately measure what it intends to assess. Thus, while high reliability is essential for trustworthiness in data, it must be coupled with validity to ensure that research findings are meaningful and applicable. In practice, ensuring reliability involves standardizing data collection procedures, training data collectors, and maintaining consistency in test environments. This framework is equally applicable in educational contexts, especially for summative assessments where accurate, bias-free, and reproducible test results are critical for fair student evaluation.

Thus, summative assessment is essential in teacher education for certifying learning and professional readiness. However, studies reveal that many assessments lack validity and reliability [15], [23], [3]. Validity ensures alignment with learning goals, covering content, construct, criterion, and face validity [2] [23]. Reliability, on the other hand, ensures consistency of results across time and raters [9]. While reliability alone does not guarantee validity, both are necessary for fair and accurate assessment. This highlights the need for validated, reliable, and curriculum-aligned tests in BTLEd programs.

Methodology

Research Design

This study used a descriptive research design to evaluate a summative assessment tool in terms of validity, reliability, and item analysis. Descriptive research is appropriate for examining conditions as they naturally occur, especially in educational settings where test quality and alignment with learning goals are assessed [27]. This method is valued for its simplicity, flexibility, and utility in diverse contexts[28].

Respondents of the Study

To ensure the quality of the developed summative assessment tool, two sets of respondents were involved: expert validators for content validity and student respondents for reliability testing.

Content Validity Respondents (Expert Validators)

Two subject matter experts evaluated the test for content validity to ensure that the assessment tool was accurate, relevant, and aligned with the intended learning outcomes of the course. The first expert had 28 years of teaching experience in professional education courses, particularly in curriculum development, and held a trainer's methodology certificate, demonstrating expertise in both instructional content and pedagogy. Her qualifications ensured a thorough review of the test items in terms of clarity, appropriateness, and curricular alignment.The second expert brought 22 years of teaching experience and served as a department chairperson, regularly reviewing faculty-developed test questionnaires. She was responsible for ensuring that test items adhered to the Table of Specifications (TOS) and appropriately reflected the cognitive levels of Bloom's Taxonomy. With substantial training both nationally and internationally in assessment and evaluation, she provided informed insights to strengthen the content validity of the instrument. Their combined expertise ensured that the summative test accurately measured the essential competencies outlined in the course Curriculum Development and Evaluation with Emphasis on TM II.

Technium Social Sciences Journal
Vol. 74, 96-112, August, 2025
ISSN: 2668-7798
www.techniumscience.com

Reliability Testing Respondents

To determine the reliability of the instrument, the summative test was administered to 41 third-year Bachelor of Technology and Livelihood Education (BTLEd) students who had previously completed the Curriculum Development and Evaluation with Emphasis on TM II course during their second year. These students were purposively selected due to their familiarity with the course content, making them well-suited to assess the clarity and consistency of the test items. Their background enabled them to provide meaningful responses, ensuring that the data collected aligned with the study's objectives.

Table 1. Demographic Profile of the Reliability Testing Respondents

Profile		Frequency	Percentage
Age	20 and below	1	2.44
	21-22	4	9.76
	23-24	32	78.04
	25-26	2	4.88
	27-28	1	2.44
	29 and above	1	2.44
Sex	Female	29	70.73
	Male	12	29.27

The data in Table 1 show that the majority of respondents (78.04%) were between the ages of 23 and 24, indicating a typical age range for third-year college students. Most participants were female (70.73%), reflecting a higher female enrollment in the BTLEd program. These demographics suggest a relatively homogenous group in terms of age and gender, which may contribute to consistent responses during the assessment process.

Development of the Summative Test

The research instrument consisted of an 80-item multiple-choice summative test, designed to assess BTLEd students' understanding of the Curriculum Development with Emphasis on Trainers' Methodology II course. The test followed the course syllabus to evaluate both content understanding and practical application of essential concepts. The test construction process followed a Table of Specifications (TOS) to achieve content validity through balanced coverage of instructional topics and cognitive domains based on Bloom's taxonomy. The weekly breakdown of issues and intended learning outcomes (ILOs) in Table 2 demonstrates direct alignment with the test items, which assess cognitive, affective, and psychomotor domains. The TOS presented in Table 3 illustrates how test items were distributed based on instructional time, topic weight, and cognitive level targets. These tables collectively show that the summative assessment was designed to match course content and learning objectives, thus validating its ability to measure student performance effectively.

Technium Social Sciences Journal
Vol. 74, 96-112, August, 2025
ISSN: 2668-7798
www.techniumscience.com

Table 2. Weekly Topics and Intended Learning Outcomes for the course Curriculum Development and Evaluation with Emphasis on TM II

Week	Hours	Topics	Intended Learning Outcome
Week 1	6 hours	**Understanding Curriculum Development and Evaluation with Emphasis of TM II** • Definition of Curriculum • Different Types of Curriculums • Curriculum Foundation, Conception and elements • Overview of Trainers' Methodology II Training Needs Assessment and Analysis	a) Explain the foundations, elements, and various types of curriculums used in education and training. b) Describe the relevance of TM II and the role of Training Needs Assessment in curriculum development. c) develop a context-specific Training Needs Assessment (TNA) instrument in a selected community, and collect empirical evidence to identify priority training needs
Week 2	9 hours	**Planning and Crafting the Curriculum** • Fundamentals of Curriculum Designing • Approaches to Curriculum Designing • Curriculum Development: Processes and Models • Process of Training Curriculum Development	a) Analyze different approaches and models of curriculum design. b) Apply the curriculum development process to design a training curriculum/Session Plan aligned with identified needs. c) Justify the selection of appropriate design models for various learning contexts.
Week 3	9 hours	**The Teacher as Curriculum Implementor and Manager** • Setting Clear Learning Outcomes • Selecting and Organizing Curriculum Content & Instructional Strategies • The Role of Technology in delivering and designing the curriculum	a) Formulate clear and measurable learning outcomes aligned with identified curriculum goals and training needs. b) Organize curriculum content and instructional strategies that support learner-centered delivery. c) Produce a training module that integrates SMART objectives, relevant content, credible resources, and appropriate technology.
Week 4	9 hours	**Curriculum Evaluation** • What, Why and How to Evaluate a Curriculum • Curriculum Evaluation Through Learning Assessment	a) Analyze the purpose, principles, and process of curriculum evaluation to enhance instructional effectiveness. b) Value the role of learning assessments and evaluation data in improving curriculum relevance and quality.

TECHNIUM SOCIAL SCIENCES JOURNAL

Technium Social Sciences Journal
Vol. 74, 96-112, August, 2025
ISSN: 2668-7798
www.techniumscience.com

			c) Construct assessment tools that align with intended training outcomes and effectively measure learner performance and skill acquisition.
Week 5	9 hours	**Outcomes-Based-Education: Basis for Enhanced Teacher Education Curriculum** • Four Essential Principles in OBE • Teaching-Learning in OBE • Assessment of Learning Outcomes in OBE • Utilization of Technology in Training	a) Explain the core principles of Outcomes-Based Education (OBE) and analyze their implications for curriculum and instruction in teacher education. b) Demonstrate commitment to learner-centered teaching by aligning instructional strategies and assessments with clearly defined learning outcomes. c) evaluate and revise a learning module that integrates OBE principles, appropriate technology, and performance-based assessment strategies.
Week 6	6 hours	**Alignment in the 21st Century Learning Environment** • Curricular landscape in the 21st Century Classroom • The 7 Rs of Quality Curriculum Material	a) Identify key features of a 21st-century curriculum. b) Apply the 7 Rs in developing curriculum materials that are responsive to learner needs. c) Align curriculum content with 21st-century skills and learning demands.

Table 3. Table of Specification for Curriculum Development and Evaluation with emphasis on TM II

TOPIC/S	No. of Hours Taught	%	No. of Test Items/ Points	COGNITIVE LEVEL			
				Knowledge %	Comprehension %	Application Analysis %	/Synthesis / Evaluation %
				20%	20%	50%	10%
Understanding Curriculum Development and Evaluation with Emphasis of TM II Developing TNA	6	12.5	10	2	2	5	1
Planning and Crafting the Curriculum Designing Session Plan	9	18.75	15	3	3	7	2
The Teacher as Curriculum Implementor and Manager Creating Learning Module	9	18.75	15	3	3	7	2

Technium Social Sciences Journal
Vol. 74, 96-112, August, 2025
ISSN: 2668-7798
www.techniumscience.com

Curriculum Evaluation Designing Assessment Task/Tool	9	18.75	15	3	3	7	2
Outcomes-Based-Education: Basis for Enhanced Teacher Education Curriculum	9	18.75	15	3	3	9	
Alignment in 21st Century Learning Environment	6	12.5	10	2	2	5	1
TOTAL:	48	100%	80	16	16	40	8

The Table of Specifications (TOS) presented in Table 3 ensures that the course content aligns with the instructional hours and cognitive levels as described by Bloom's Taxonomy [27]. The 80 test items distributed across the topics corresponded to the instructional time duration. The content received 15 test items (18.75%) for each topic, spanning nine hours, and 10 items (12.5%) for each topic taught during six hours, to ensure balanced content representation.

The test design focuses on higher-order thinking, comprising 40 items (50%) for Application/Analysis, 16 items (20%) for Knowledge and Comprehension, and eight items (10%) for Synthesis/Evaluation. The assessment distribution aligns with the course objectives to evaluate both conceptual understanding and practical application, as these skills are crucial for teacher education programs, according to [27]. Higher-order thinking emphasis enables students to evaluate curriculum principles both critically and practically in actual teaching situations [29]. The initial construction of 160 items exceeded the 80 required items, providing sufficient space for improvement. Researchers consider building an extensive initial item collection a fundamental practice in developing questionnaires and tests. DeVellis [30] and other researchers emphasize that construct operationalization involves creating a comprehensive list of potential items before selecting the most representative and effective ones. No statistical method exists to fix poorly constructed or missing items according to [30].

Experts suggest building the initial item pool to contain 2 to 4 times more items than the planned final count, enabling item improvement and elimination through statistical analysis [30]. This methodology helps select high-performing items while enhancing internal consistency reliability, resulting in stronger instrument validity [26]. Early content redundancy functions as a beneficial aspect because it helps maintain the most reliable items throughout the development of the final test.

The strategy ensures instrument accuracy and fairness by retaining statistically sound and pedagogically valid items in the final version. The developers of the assessment method created each test question to evaluate both memory and analytical skills, as well as practical application competencies. Two subject experts reviewed the draft test to assess the clarity, relevance, and alignment of the items. The necessary changes to the instrument were informed by feedback, which improved its content validity, as noted by [25].

The test administration employed a paper-and-pencil method, designed to replicate standard testing situations. Research confirms that multiple-choice assessments effectively

Technium Social Sciences Journal
Vol. 74, 96-112, August, 2025
ISSN: 2668-7798
www.techniumscience.com

measure retention and comprehension when they are correctly aligned with learning objectives [25]. This organized method led to the development of a dependable instrument that accurately measures student learning outcomes.

Data Collection Procedure

The data collection procedure involved validation, and administration of the 160-item multiple-choice summative assessment aligned with the *Curriculum Development and Evaluation with Emphasis on TM II course*. Prior to its administration, content validation was carried out using the course syllabus and a structured Table of Specifications (TOS) to ensure that the test items reflected the intended learning outcomes and were appropriately distributed across Bloom's cognitive levels [28]. The TOS is widely recognized as an effective tool for enhancing content validity by aligning assessment items with instructional goals and cognitive demands [28]. Following the validation process, the assessment was administered to BTLEd students who had completed the course and were deemed capable of providing relevant data based on their familiarity of the content.

Data Analysis

After the administration of the assessment, the test results were subjected to statistical analysis to determine the psychometric soundness of the instrument. Internal consistency reliability was measured using the Kuder-Richardson Formula 20 (KR-20), which is appropriate for assessments with dichotomous scoring formats such as multiple-choice tests [31]. In addition, item analysis was performed to examine each test item's difficulty and discrimination indices, key indicators used to assess the effectiveness of items in differentiating between high- and low-performing students [31] these analytical procedures ensured that the assessment tool was both valid and reliable, supporting its intended use as a summative evaluation measure in the context of teacher education.

Results and discussion

Validity of Summative Test

The summative test for Curriculum Development and Evaluation with emphasis on TM II, underwent a thorough content validation process to verify its alignment with learning objectives, the course syllabus, and the Table of Specifications (TOS). Two subject matter experts who specialized in curriculum development and assessment evaluated test items in depth. The evaluation examined how well test items presented themselves and their connection to Bloom's Taxonomy cognitive levels. The experts made multiple revisions to the test items, improving phrasing clarity and distractor plausibility, while also matching cognitive demands to the course learning objectives.

Expert 1 recommended changes to 62 items due to grammatical errors and unclear wording that impacted understanding. Expert 2 suggested improvements to 58 additional items, focusing on making the language more precise and maintaining consistent terminology. The experts provided different levels of feedback, but their suggestions supported each other by emphasizing both the quality of items and their alignment with learning objectives. T

The implementation of systematic expert feedback led to improved test items, thereby enhancing the instrument's content validity. The validation process confirmed that the final test effectively combined both instructional quality and psychometric strength to measure student proficiency in curriculum development and evaluation competencies.

Technium Social Sciences Journal
Vol. 74, 96-112, August, 2025
ISSN: 2668-7798
www.techniumscience.com

Reliability of Summative Test

Table 4 presents the reliability analysis of the developed summative test using the Kuder-Richardson Formula 20 (KR-20). This analysis provides evidence of the internal consistency of the assessment instrument administered to 41 BTLEd students.

Table 4. Reliability Result of Summative Test

No. of items	No. of Takers	Mean for Test	Standard Deviation for Test	KR-20 Value	Interpretation
160	41	64.88	17.69	0.94	Very High Reliability

Table 4 reveals a KR-20 value of 0.94, indicating that the assessment meets psychometric standards for internal consistency at a very high level. Test items show reliable performance when their coefficient exceeds 0.70, with values above 0.80 indicating high reliability [32].

The assessment method achieves good reliability because it demonstrates strong performance in measuring learning outcomes. The instrument produces scores ranging from 64.88 to 17.69 in mean and standard deviation, indicating students demonstrate various levels of understanding. The results demonstrate that the summative test is a reliable and effective tool for assessing student achievement.

A high internal consistency level in test development becomes vital because it guarantees both accuracy and dependability of assessment tools [32]. A reliability coefficient below acceptable standards would require item revisions through question clarification and difficulty calibration to enhance measurement precision and fairness.

The expert validation process, along with the structured TOS development, ensured strong content validity by linking both the summative test items to the TOS and the course syllabus. The assessment of theoretical knowledge and practical competencies in teacher education received a fair and consistent evaluation through this process, which also demonstrated high reliability.

Item difficulty of Summative Test

Table 5 displays the test items along with their item difficulty index (*IDI*), which indicates the percentage of students who correctly answered each item [32]. The analysis serves to identify the difficulty level of items, which range from very easy to very difficult. The assessment of student performance levels requires a test that contains diverse item complexities, which item difficulty analysis helps to achieve [30]. Hartati and Yogi [33] suggest using a 1:2:1 distribution pattern, where 25% of items should be easy, 50% should be moderate, and 25% should be difficult, to maintain both fairness and discrimination power. The approach enhances discrimination performance and enables the generation of accurate and reliable results.

Technium Social Sciences Journal
Vol. 74, 96-112, August, 2025
ISSN: 2668-7798
www.techniumscience.com

Table 5. *Distribution of items based on Difficulty*

Item Difficulty Index	Difficulty level	Number of Questions	Percentage
0.86 - 100.00	Very Easy	22	13%
0.66 – 0.85	Easy	49	30%
0.46 – 0.65	Moderate	55	34%
0.26 – 0.45	Difficult	11	6%
00.00 – 0.25	Very Difficult	23	14%

The distribution of test items, based on their difficulty indices, is presented in Table 5. The summative assessment and the evaluation of student learning across different proficiency levels benefit from 55 items (34%) that fall within the optimal difficulty range of (*IDI* = 0.46–0.65) (Brown, 2004). The test items were calibrated correctly to measure average student comprehension since they comprised nearly half of the total assessment content. The test contained 49 items (30%) that were classified as easy (*IDI* = 0.66–0.85) and 22 items (13%) that were categorized as very easy (*IDI* = 0.86–1.00). The easy and very easy items made up 43% of the total test content, which indicates that most students could access these sections to support inclusive and fair assessment practices.

The assessment contained 11 items (6%) that were classified as difficult (*IDI* = 0.26–0.45) and 23 items (14%) that were very difficult (*IDI* = 0.00–0.25), which together made up 20% of the total items. The assessment contains a small number of challenging items, which may be suitable for specific assessment goals. According to Ntumi et al. [32], end-of-course summative tests should include moderate-level items because they provide effective differentiation of student performance across various skill levels. The use of more challenging items usually occurs for diagnostic or selection assessments [32]. The assessment tool demonstrates a balanced design because it focuses on moderate and accessible items, which enhance both fairness and validity. The evaluation contains a limited number of challenging items, which helps students feel more confident during testing while maintaining effective measurement of their learning achievements.

Item Discrimination of Summative Test

Discriminating power refers to a test item's ability to distinguish between high- and low-performing students. Items with high discrimination are answered correctly by those with higher scores and incorrectly by those with lower scores, reflecting actual differences in ability [32]. The discrimination index, ranging from -1.0 to 1.0, indicates item quality, where higher positive values show better discrimination [32]. Items near or below zero may require revision.

Table 6 shows the distribution of items based on their discrimination indices, helping identify which items are strong, acceptable, or weak in distinguishing student performance. This analysis is vital to ensure fairness and validity in assessment.

Technium Social Sciences Journal
Vol. 74, 96-112, August, 2025
ISSN: 2668-7798
www.techniumscience.com

Table 6. Distribution of items based on Discriminating Power

Discriminating Power	Discriminating Power Interpretation	Number of Questions	Percentage
0.40 and above	Very Good Item; accept	65	40%
0.30-0.39	Reasonably good but subject to improvement	37	23%
0.20-0.29	Marginal items usually need and subject to improvement	20	13%
Below 0.19	Poor items to be rejected	38	24%

Table 6 presents the analysis of the discrimination indices for the 160 test items. The test items included 65 very good items (40%) that demonstrated discrimination indices exceeding 0.40, effectively separating high- and low-performing students. The items fall under the acceptable category and need no significant modification [31]. The 0.30–0.39 range encompasses 37 items that are classified as reasonably good but require minor adjustments to enhance their discriminative ability.

The 0.20–0.29 range contains 20 items, which are classified as marginal and need revision to enhance their discriminative ability. The 38 items (24%) that scored below 0.19 are considered poor discriminators because they fail to separate knowledgeable from less knowledgeable students, thus requiring rejection, as per [31] and [32].

A high discrimination index, as indicated by [32], suggests that an item effectively measures student understanding, making it crucial for valid assessments. Test reliability and fairness suffer when items have low or negative discrimination values, as they create confusion among students and fail to align with learning objectives. The results demonstrate that most items performed well; however, a few require improvement or removal to enhance the assessment tool's overall quality and effectiveness.

The test quality becomes evident when analyzing the results together with the item difficulty data. The majority of items fell into moderate or easy categories, yet 38 items were rejected because they failed to demonstrate discrimination in student performance. The results reveal why both difficulty and discrimination indices serve as essential validation tools for tests. The test becomes more accurate and learning outcomes are aligned by removing poorly performing items.

Conclusion

As competency-based learning becomes increasingly integral to teacher education, the need for highly valid and reliable assessments that are closely aligned with instruction is essential for evaluating both student readiness and instructional effectiveness. This study aimed to develop a psychometrically valid and reliable summative test designed explicitly for the Curriculum Development and Evaluation course, with an emphasis on TM II, for Bachelor of Technology and Livelihood Education (BTLEd) students. The assessment was developed to meet quality assessment standards by incorporating fairness principles, curriculum alignment, and diagnostic utility elements, which are fundamental to quality assessment practice.

The instrument achieved high content validity through its Table of Specifications (TOS) and expert reviews, as well as its detailed alignment with the course syllabus. The assessment demonstrated exceptional internal consistency, with a KR-20 coefficient of 0.94,

which confirms its reliability. The item analysis revealed that the test items maintained moderate difficulty levels and strong discrimination indices, making them effective at separating student performance while providing helpful feedback. The assessment eliminated 38 items from the initial 160 due to their unsatisfactory statistical results, resulting in 122 valid test items. A final 80-item instrument was carefully chosen through established psychometric standards and content relevance to maintain both validity and reliability. The research makes an essential contribution to teacher education by presenting a systematic approach to test development that combines expert opinions with statistical methods and instructional alignment. Although the study focused exclusively on one course within a specific institution, the process demonstrates how to enhance assessment design in educational contexts with similar characteristics. Additional research should focus on testing the instrument across multiple institutions, subjects, and student groups to broaden its applicability and increase its effects.

A well-validated summative assessment serves as both an essential element of instructional quality and a tool for evaluating educational equity and learner competence. Educators should prioritize assessments with strong psychometric properties combined with context-specific relevance to support the development of competent teachers who can meet modern academic requirements.

Ethical Considerations

In the conduct of this study, the researcher adhered to established ethical standards to protect the rights and welfare of all participants. Prior to data collection, informed and voluntary consent was obtained from all respondents, and the purpose and objectives of the research were clearly explained to ensure transparency and mutual understanding. Formal permission was also secured from the immediate head of the institution where the study was conducted. To safeguard participant privacy, strict confidentiality measures were observed. All personal information was anonymized and securely stored to prevent unauthorized access or disclosure.

Additionally, all research materials, including test questionnaires and informed consent forms, were carefully constructed and reviewed to ensure clarity and to eliminate any form of bias, offensive content, or discriminatory language. These measures were implemented to uphold the integrity of the research and ensure ethical compliance throughout the study process.

References

[1] Earle, S. (2020). Balancing the demands of validity and reliability in practice: Case study of a changing system of primary science summative assessment. *London review of education* https://doi.org/10.14324/LRE.18.2.06

[2] Kibble, J. D. (2017). Best practices in summative assessment. *Advances in Physiology Education*, *41*(1), 110119. https://doi.org/10.1152/advan.00116.2016

[3] Dogara, G., Kamin, Y. Bin, & Saud, M. S. Bin. (2020). The Impact of Assessment Techniques on
the Relationship between Work-Based Learning and Teamwork Skills Development. *IEEE Access*,
8, 59715–59722. https://doi.org/10.1109/ACCESS.2020.2983487

[4] Siti, S. R., Rasul, M. S., Mohammad Yasin, R., & Hashim, H. U. (2023). Identifying and Validating Vocational Skills Domains and Indicators in Classroom Assessment Practices in TVET.
Sustainability 2023, Vol. 15, Page 5195, *15*(6), 5195. https://doi.org/10.3390/SU15065195

[5] Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability,* 21(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

[6] Pintrich, P. R. (2002). The role of metacognitive knowledge in learning, teaching, and assessing. *Theory into Practice,* 41(4), 219–225. https://doi.org/10.1207/s15430421tip4104_3

[7] Paul Black , Christine Harrison , Jeremy Hodgen , Bethan Marshall & Natasha Serret (2010) Validity in teachers' summative assessments, Assessment in Education: *Principles,*
Policy & Practice, 17:2, 215-232, https://doi.org/10.1080/09695941003696016

[8] Orongan, R. C., Manual, A. A., Orongan, M. J. Q., Gamutan, L. V., & Vegafria, J. C. (2024). Analysis on Teachers' Summative Assessment in Basic Secondary Education (pp. 443–457). https://doi.org/10.2991/978-94-6463-554-6_36

[9] Gupta, K. (2023). Validity and Reliability of Students' Assessment: Case for Recognition as a Unified Concept of Valid Reliability. *International Journal of Applied & Basic Medical Research*, *13*(3), 129–132. https://doi.org/10.4103/ijabmr.ijabmr_382_23

[10] French, S., Dickerson, A., & Mulder, R. A. (2024). A review of the benefits and drawbacks of high-stakes final examinations in higher education. In *Higher Education* (Vol. 88, Issue 3, pp. 893–918). Springer Science and Business Media B.V. https://doi.org/10.1007/s10734-023-01148-z

[11] Banerjee, S., Rao, N. J., Ramanathan, C., & Ramesh, V. (2014). Design of valid summative assessment instruments in formal higher education programs. *Proceedings - IEEE 6th International Conference on Technology for Education, T4E 2014*, 76–79. https://doi.org/10.1109/T4E.2014.52

[12] Newstead, S. (2002). Examining the examiners: Why are we so bad at assessing students? In *Psychology Learning and Teaching* (Vol. 2, Issue 2). https://doi.org/https://doi.org/10.2304/plat.2002.2.2.70

[13] Esmael, N. T., & Rabut, J.F. (2025). Evaluating teacher-made mathematics test and teachers' competency in assessment. *International Journal of Research and Innovation in Social Science (IJRISS)*, 6(8), 1062–1068. https://dx.doi.org/10.47772/IJRISS.2025.90400113

[14] Apolinaria Daquioag-Andres, M. C. (2023). Establishing quality summative assessment for pre-service elementary teachers: A psychometric approach. *Profile: Issues in Teachers' Professional Development*, 24(2), 217–231. http://dx.doi.org/10.47750/jett.2023.14.03.002

[15] Glenn M. Gambi, Ph.D. & OLGA C. ALONSABE, Ph.D., 2024. "**Examining and Validating Summative Tests Used in Competency-Based Assessment for TESDA Technology Institutions (TTI): Basis for Training Design**," International Journal of Research and Innovation in Social Science, International Journal of Research and Innovation in Social Science (IJRISS), vol. 8(7), pages 936-949, July.

[16] Napanoy,J.B., Peckley, M.K., (2020). Assessment Literacy of Public Elementary School Teachers in the Indigenous Communities in Northern Philippines. *Universal Journal of Educational Research*, 8(11B), 5693 - 5703. **https://doi.org**/10.13189/ujer.2020.082203

[17] Austria-Cruz, M. C. A. (2019). Academic Stress and coping Strategies of Filipino College Students in private and public universities in Central Luzon. *International Journal of Advanced Engineering, Management and Science*, 5(11), 603–607. https://doi.org/10.22161/ijaems.511.6

[18] Pastore S (2023) Teacher assessment literacy: a systematic review. Front. Educ. 8:1217167. https://doi.org/10.3389/feduc.2023.1217167

[19] Schildkamp, K., van der Kleij, F. M., Heitink, M. C., Kippers, W. B., & Veldkamp, B. P. (2020). Formative assessment: A systematic review of critical teacher prerequisites for classroom practice. *International Journal of Educational Research, 103*, 101602. https://doi.org/10.1016/j.ijer.2020.101602

[20] Tienson-Tseng, H. L. (2019). Best practices in summative assessment. In B. A. Z. Shakhashiri, S. A. Gellman, & A. B. Ellis (Eds.), *Biochemistry education: From theory to practice* (pp. 219–243). ACS Symposium Series, Vol. 1337. https://doi.org/10.1021/bk-2019-1337.ch010

[21] Jauhari, A., Dwi Putra Negara, Y., Arif Efendi, F., Ayu Mufarroha, F., Rosa Anamisa, D., & Basuki, A. (2023). Determining Student Achievements using Multicriteria Approach with Electre Method. In *Technium: Romanian Journal of Applied Sciences and Technology* (Vol. 16). https://doi.org/10.47577/technium.v16i.9984

[22] Setiabudi, A., Mulyadi, M., & Puspita, H. (2019). An Analysis of Validity and Reliability of a Teacher-Made Test. *Journal of English Education and Teaching*, *3*(4), 522–532. https://doi.org/10.33369/jeet.3.4.522-532

[23] Rizky Hilaldy. [26] Rizky Hilaldy. (2021). Content Validity of English Summative Test. *Journal of Language Learning and Research (JOLLAR)*, *4*(1), 37–41. https://doi.org/10.22236/jollar.v4i1.7708

[24] Puspitasari, E.(2022). An Analysis of Content Validity on Summative Test for the 12TH Grade Students. *Journal of English Pedagogy and Applied Linguistics, 2(2),120-131.* https://doi.org/10.32627/jepal.v2i2.420

[25] Suhaini, M., Ahmad, A., & Bohari, N. M. (2021). Assessments on vocational knowledge and skills: A content validity analysis. In *European Journal of Educational Research* (Vol. 10, Issue 3, pp. 1529–1540). Eurasian Society of Educational Research. https://doi.org/10.12973/EU-JER.10.3.1529

[26] Ahmed, I., & Ishtiaq, S. (2021). Reliability and validity: Importance in Medical Research. In *Journal of the Pakistan Medical Association* (Vol. 71, Issue 10, pp. 2401–2406). Pakistan Medical Association. https://doi.org/10.47391/JPMA.06-861

[27] Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approach* (4th ed.). SAGE. Retrieved from https://spada.uns.ac.id/pluginfile.php/510378/mod_resource/content/1/creswell.pdf

[28] DiDonato-Barnes, N., Fives, H., & Krause, E. S. (2014). Using a Table of Specifications to improve teacher-constructed traditional tests: An experimental design. *Assessment in Education: Principles, Policy and Practice*, *21*(1), 90–108. https://doi.org/10.1080/0969594X.2013.808173

[29] Anderson, L. W., & Krathwol, D. R. (2001). *A Taxonomy of Learning, Teaching, and Assessing: A revision of Bloom's Taxonomy of Educational Objectives*.

[30] DeVellis, R. F. (2017). *Scale development: theory and applications* (Olivia Weber-Stenis, Ed.; Fourth edition). SAGE.

[31] Mitra, N. K., Ponnudurai, G., Haleagrahara, N., & Judson, J. (2009). The Levels of Difficulty and Discrimination Indices in Type A Multiple Choice Questions of Pre-Clinical Semester 1 Multidisciplinary Summative Tests. *International E-Journal of Science, Medicine & Education*, *3*(1), 2–7. https://doi.org/10.56026/imu.3.1.2

[32] Ntumi, Simon, et al. Estimating the Psychometric Properties (Item Difficulty, Discrimination and Reliability Indices) of Test Items Using Kuder-Richardson Approach (KR-20). *Shanlax International Journal of Education*, vol. 11, no. 3, 2023, pp. 18–28.

Technium Social Sciences Journal
Vol. 74, 96-112, August, 2025
ISSN: 2668-7798
www.techniumscience.com

[33] Hartati, N., &Yogi, H. P. S. (2019). Item Analysis for a Better-Quality Test. *English Language in Focus* (ELIF), 2(1), 59–70. https://doi.org/10.24853/elif.2.1.59-70