International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Enhancing Students' Understanding of Validity and Reliability through the CREATE Approach (Critically Reflect, Explore, Analyze, Test, and Evaluate)

Nerlie C. Amodo[1], Dellen May M. Casilao[2], Danna D. Hadjali[3], Haneen F. Lingasa[4], Harold Kim Malubay[5], John Paul Ramon[6], Hadiya T. Taduman[7]

[1,2,3,4,5,6,7]Student, Bachelor of Technology and Livelihood Education major in Home Economics, Faculty of Teacher Education, Davao Oriental State University, Mati City, Davao Oriental, Philippines

Abstract

This classroom-based action research examined the effectiveness of the CREATE Approach (Critically Reflect, Explore, Analyze, Test, and Evaluate) in improving second-year Bachelor of Technology and Livelihood Education major in Home Economics (BTLEd-HE) students' understanding and application of validity and reliability in assessment. Existing studies show that pre-service teachers often have difficulty translating theoretical knowledge of assessment into classroom practice, and limited experiential and contextualized learning has been identified as a major challenge. The study addressed this gap through a structured, experiential and modular intervention implemented over four weeks, targeting the memorization of definitions, difficulty distinguishing validity from reliability, and limited exposure to time-bounded, scenario-based tasks. Nineteen of 26 students who had difficulty with these concepts, identified through diagnostic assessment and classroom performance, participated. An embedded mixed-methods design was used; data were gathered through a 30-item pretest and posttest, rubric-scored error analysis and peer testing tasks, and reflection journals. The mean score increased from 16.79 (SD = 3.87) in the pretest to 25.37 (SD = 1.61) in the posttest, and all 19 students reached the 75% passing standard after the intervention, compared with none before it. A paired-sample t-test showed a statistically significant difference, $t(18) = 9.37$, $p < 0.001$, with a large effect size (Cohen's $d = 2.15$). All students reached the Excellent level in the error analysis and peer testing validation tasks, and the reflection journals showed clearer conceptual distinctions, more critical thinking, greater confidence and stronger peer collaboration. The continued use of experiential, scenario-based activities, collaborative peer evaluation and reflection journals is recommended, together with individual guidance, varied illustrative examples and flexible support for latecomers and students with short attention spans. The CREATE Approach offers a practical and replicable model for developing assessment literacy in pre-service teacher education.

Keywords: Assessment Literacy, CREATE Approach, Experiential Learning, Validity and Reliability, Pre-Service Teachers, Action Research, Simulation

1. Introduction

Assessment literacy is a primary competence that every pre-service teacher has to develop. The Commission on Higher Education (CHED), through Memorandum Order No. 78, Series of 2017, requires graduates of the Bachelor of Technology and Livelihood Education program to construct and administer assessments that are valid and reliable [1]. This requirement is also reflected in the Philippine Professional Standards for Teachers (PPST), which state that teachers must be capable of designing, selecting, organizing and using a variety of assessment strategies to assess learning [2]. Mastery of these concepts is essential because validity ensures that assessments measure the intended learning outcomes, while reliability ensures that the results are consistent, which makes assessment literacy a priority issue in teacher education.

Many pre-service teachers find it difficult to apply assessment knowledge in practice. Even when they understand assessment concepts in theory, they have the hardest time transferring that knowledge into real classroom situations. A systematic review by Pastore (2023) notes that pre-service teachers struggle to apply assessment principles effectively, mainly because of the transition from theory to practice; limited practice opportunities and lecture-dominated instruction can lead to shallow understanding, poor alignment of assessments with learning objectives and indecision in making assessment judgments [3]. In Indonesia, a study of English as a Foreign Language (EFL) teachers found that although the teachers understood assessment principles well, their competence in designing and implementing classroom assessments remained at a medium level [4]. In Ghana, teachers' conceptions of validity and reliability were found to influence how they design classroom assessments [5].

At Davao Oriental State University (DOrSU), 26 second-year BTLEd-HE students were enrolled in EDUC 107: Assessment in Learning 1. Diagnostic results and classroom performance showed that 19 of them (73%) persistently had difficulty telling validity and reliability apart and applying the two concepts to assessment tasks. The difficulty was most apparent in situational and timed assessments, in which students mixed up the two concepts, misjudged test items and gave weak justifications, which reflected limited conceptual understanding and little experience in constructing and analyzing tests. Further probing through scenario-based activities, a student survey and a Why-Why analysis revealed root causes such as a lack of contextualized activities, limited interaction, little experiential practice and a lack of simulated assessment tasks.

Strengthening this competence is important because pre-service teachers who cannot apply validity and reliability may administer inconsistent, unfair and inaccurate assessments in their future classrooms, which can harm learners' performance and make sound instructional decisions difficult. To address this need, the researchers developed the CREATE Approach (Critically Reflect, Explore, Analyze, Test, and Evaluate), an experiential learning model grounded in Experiential Learning Theory, Constructivism, Social Learning Theory, Cognitive Load Theory and Formative Assessment Theory. The model guides students through a structured cycle of reflection, exploration, error analysis, assessment design and evaluation, combining scenario-based tasks, collaborative analysis, mini-test construction, peer validation and reflective journaling. All materials and tools used in the intervention were validated by subject-matter experts before implementation to ensure their appropriateness, accuracy and alignment with assessment principles.

1.1 Problem Statement

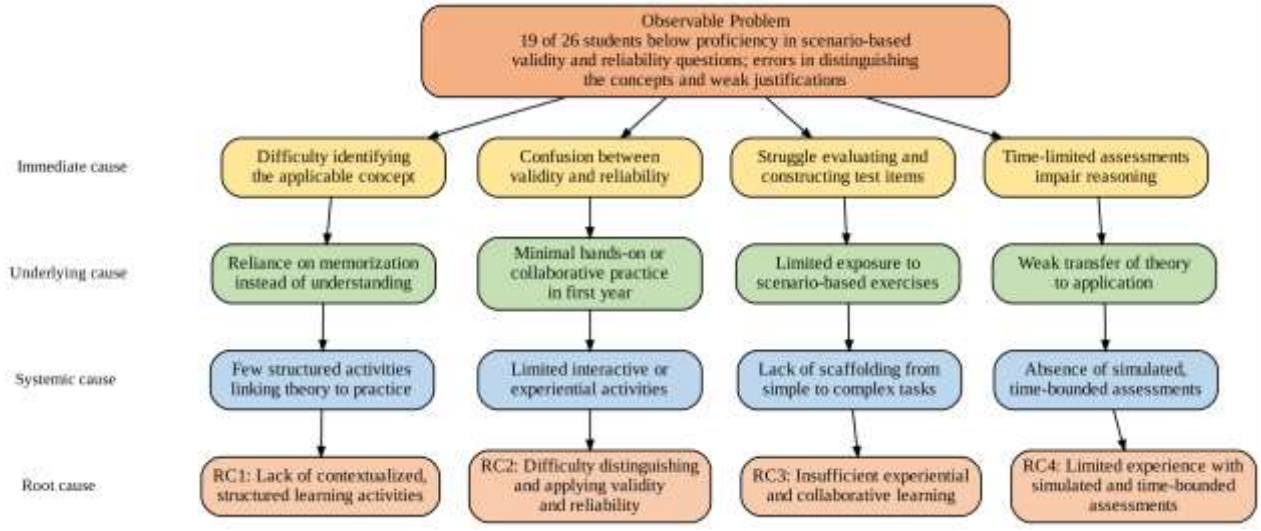
BTLEd-HE second-year students at DOrSU Main Campus had difficulty understanding and applying the principles of validity and reliability in EDUC 107: Assessment in Learning 1. Classroom observation and

a survey showed that 19 of 26 students could not identify or apply these principles when analyzing test items and situational questions. In the survey of these 19 students, 15 (78.95%) reported difficulty analyzing situational questions about validity and reliability, five (26.32%) reported difficulty distinguishing the two concepts, and three (15.79%) reported time pressure when answering situational questions. Although students could memorize the definitions, many could not distinguish and apply the principles when presented with actual assessment situations, could not connect a scenario with the correct concept, and felt confused when applying the concepts to real-life examples. These difficulties appear to stem from prior instruction that emphasized memorization of definitions rather than application. The problem is classroom-based and within the teacher's control, and it must be addressed because the ability to apply validity and reliability is a core skill that BTLEd-HE students need to construct fair, consistent and effective assessment tools in their future teaching careers.

1.2 Root Cause Analysis

A root cause analysis using the Why-Why technique traced the problem from its observable symptoms to its root causes (Figure 1). The observable problem was that 19 of 26 students scored below proficiency in scenario-based questions on validity and reliability, confused the two concepts and gave weak justifications. The immediate causes were difficulty identifying the applicable concept in a scenario, confusion between validity and reliability, difficulty evaluating and constructing test items, and time-limited assessments that impaired reasoning. At the underlying level, students relied on memorization rather than understanding, had minimal hands-on or collaborative practice in their first year, had limited exposure to scenario-based exercises, and showed weak transfer of theory to application. At the systemic level, earlier coursework offered few structured activities linking assessment theory to practice, limited interactive or experiential activities, little scaffolding from simple to complex tasks, and no simulated or time-bounded assessments.

Figure 1: Why-Why Diagram Tracing the Causal Chain from the Observable Problem to the Root Causes



The analysis showed that the performance gaps stem from specific instructional conditions rather than from a lack of ability among the students. Four root causes within the teacher's control were identified:

1. Root Cause 1 (RC1): Lack of contextualized and structured activities. Students rarely had the chance to connect assessment theory to real tasks, which limited their ability to understand and analyze scenario-based questions.
2. Root Cause 2 (RC2): Difficulty distinguishing and applying validity and reliability. Because of limited exposure to interactive, guided or hands-on exercises in previous courses, students often mixed up the two concepts and relied on memorized definitions.
3. Root Cause 3 (RC3): Insufficient experiential and collaborative learning. During their first year, students had few opportunities for reflective work, peer collaboration or project-based activities, which made it harder to internalize assessment principles.
4. Root Cause 4 (RC4): Limited experience with simulated and time-bounded assessments. Students were not used to answering scenario-based questions under time pressure and therefore fell back on memorization instead of contextual reasoning.

Each root cause was checked against evidence and assessed for controllability, as summarized in Table 1.

Table 1: Root Cause Validation Matrix

Probable Root Cause	Validation	Evidence Source	Controllability	Conclusion	Priority
RC1: Lack of contextualized and structured activities connecting theory to practical tasks	True	Student survey, observation	Within control	Valid	High
RC2: Difficulty distinguishing and applying validity and reliability; reliance on memorized definitions	True	Student survey, interviews	Within control	Valid	High
RC3: Insufficient experiential and collaborative learning	True	Observation, reflection journals	Within control	Valid	Medium
RC4: Limited experience with simulated and time-bounded assessments	True	Student survey	Within control	Valid	High

1.3 Conceptual Framework

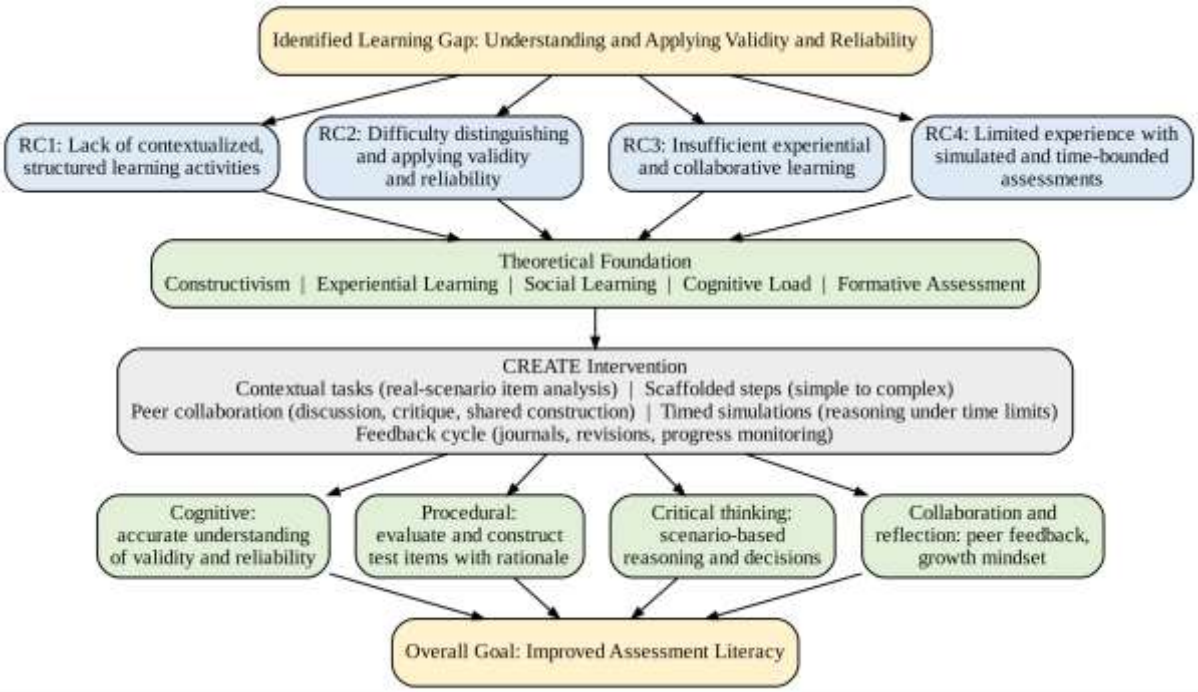
The conceptual framework is based on five educational theories that inform the design of the intervention. Constructivist Learning Theory holds that students build knowledge by connecting new information to prior experience, which helps them link prior knowledge to real-world assessment scenarios [6]. Experiential Learning Theory explains learning through experience, reflection and application, and guides the practical activities in which students evaluate test items and construct mini-tests [7]. Social Learning Theory explains how learning develops through observation, modeling and interaction with peers, which underpins the collaborative group work and feedback-driven discussions in CREATE [8]. Cognitive Load

Theory emphasizes managing cognitive load by sequencing activities from simple recall to complex scenario analysis [9]. Formative Assessment Theory emphasizes ongoing feedback and monitoring of progress, which is reflected in the reflection journals, peer review sheets and revision cycles of CREATE [10]. Table 2 maps these theories to the intervention, and Figure 2 presents the framework.

Table 2: Learning Theories Mapped to the CREATE Intervention

Learning Theory	Application in CREATE	Expected Learner Outcomes
Constructivism	Activation of prior knowledge; real-life assessment scenarios	Meaning-making and contextual understanding
Experiential Learning	Hands-on test construction; analysis and reflection cycles	Deeper mastery through doing
Social Learning	Peer-assisted collaboration and group critique	Improved reasoning, feedback and shared learning
Cognitive Load	Tasks arranged from simple to complex with guided steps	Reduced overload and stronger conceptual processing
Formative Assessment	Continuous feedback, journal reflection and revision cycles	Measurable growth and metacognitive awareness

Figure 2: Conceptual Framework of the Study



1.4 Research Questions and Hypothesis

The central question of this action research is: How does the implementation of the CREATE Approach enhance second-year BTLEd-HE students' understanding and application of validity and reliability in assessment? Specifically, the study sought to answer the following:

1. What is the baseline knowledge of the pre-service teachers regarding validity and reliability?

2. What challenges do learners experience when interpreting or analyzing situational questions involving validity and reliability?
3. How does peer review enhance the quality and effectiveness of the CREATE Approach?
4. To what extent does the CREATE Approach improve students' ability to correctly analyze situational questions involving validity and reliability?

For Research Question 4, the null hypothesis tested at the 0.05 level of significance was: There is no significant difference between the pretest and posttest scores of the students.

1.5 Scope, Delimitation and Limitation

The study covered 19 second-year BTLEd-HE students (73% of the class) who had difficulty distinguishing, understanding and applying validity and reliability. The intervention included scenario exercises, error analysis, test construction, peer review, self-reflection and a final exhibition of outputs. It lasted four weeks, with two 40-minute sessions per week (5 hours and 20 minutes in total), integrated into the regular EDUC 107 class schedule as additional learning sessions that did not replace regular instruction. The study examined the CREATE Approach as an integrated teaching method and did not examine the effects of its separate stages, other teaching methods, long-term teaching performance or achievement outside the course. The study involved one class in one institution, so its results cannot be generalized widely. Because a single group was tested once before and once after the intervention, cause-and-effect conclusions are limited and external factors may have influenced the results. The researchers acted as both facilitators and observers, which may have introduced bias; to reduce this risk, evidence from several sources was considered, including the pretest and posttest, rubric-scored activities, peer reviews and reflection journals.

1.6 Significance of the Study

For students, the intervention provides a structured and practical learning process that improves their understanding of assessment concepts and develops critical thinking, problem-solving and collaboration skills needed in their future careers as teachers. For teachers, it offers a flexible model that integrates experiential and reflective techniques into the teaching of assessment. For schools and administrators, it presents a low-cost intervention that can be adopted within existing courses to strengthen the assessment literacy of future teachers. For the academic community, it contributes to teacher education literature by showing how hands-on, collaborative and self-evaluation methods can develop assessment skills. For the broader community, it supports the preparation of teachers who can construct valid, reliable and fair assessments, leading to more accurate evaluation of learning and better instructional decisions.

2. Methodology

2.1 Research Design

An embedded mixed-methods action research design was used. In an embedded design, one type of data is nested within a larger design dominated by the other, providing supporting insights that the primary method alone cannot capture [11]. In this study, the quantitative strand was primary and followed a one-group pretest–posttest structure, while qualitative data from students' reflection journals were embedded to explain how the CREATE Approach shaped students' understanding and application of validity and reliability.

2.2 Participants

The participants were second-year BTLEd-HE students of DOrSU enrolled in EDUC 107: Assessment in Learning 1. Of the 26 students in the class, 19 (73%) were found through the diagnostic assessment and

classroom performance to have difficulty differentiating and applying validity and reliability, and these 19 students (six male and 13 female) were purposively selected as participants because they represented the learning gap that the CREATE Approach was designed to address. Their pretest scores in Week 1 served as the baseline. The intervention was implemented in close coordination with the course instructor and within the class schedule. Participant codes (S1 to S19) were used, and no personal information was included in documentation or reporting.

2.3 Research Instruments

The pretest and posttest used the same 30-item paper-and-pencil test on validity and reliability, consisting of 15 multiple-choice items and 15 situational identification items in which students classified each scenario as a validity or a reliability issue. Each correct answer earned one point. The activity outputs (initial reflection notes, Concept Summary Chart, Error Analysis Chart, Framework for Good Test Design, Assessment Design Challenge, Peer Testing Validation, reflection journal and exhibition) were scored with analytic rubrics of four criteria, each rated from 1 to 5 points, for a total of 20 points. The rubric scoring interpretation was 17–20 Excellent, 13–16 Proficient, 9–12 Satisfactory, 5–8 Developing and 1–4 Beginning. The test, the rubrics and the reflection journal prompts were content-validated by three subject-matter experts. Table 3 summarizes the data sources for each research question.

Table 3: Data Collection Schedule and Methods Matrix

Research Question	Data Source and Instrument	Timeline	Analysis Method
1. Baseline knowledge of validity and reliability	Pretest; initial reflection notes	Week 1 (Critically Reflect & Explore)	Numerical scoring of correct answers; thematic analysis
2. Challenges in analyzing situational questions	Problem-identification survey; Error Analysis Chart; reflection journals	Before the intervention; Week 2 (Examine & Analyze)	Frequency and percentage; rubric scoring; thematic analysis
3. Contribution of peer review	Peer Testing Validation rubric; reflection journals	Week 3 (Apply & Test)	Rubric scoring; descriptive statistics; thematic analysis
4. Improvement in analyzing situational questions	Posttest; reflection journals	Week 4 (Evaluate & Exhibit)	Descriptive statistics; paired-sample t-test; Cohen's d; thematic analysis

2.4 Data Analysis

Pretest and posttest scores were analyzed with descriptive statistics (mean, standard deviation, frequency and percentage) to describe students' performance before and after the intervention, and a paired-sample t-test at the 0.05 level of significance was used to determine whether the increase in scores was statistically significant. Analyses were performed in Microsoft Excel. Raw scores were converted to percentages and interpreted with a criterion-referenced scale adapted from Munawaroh et al. (2022) [12]: 0%–49% Very

Poor, 50%–69% Poor, 70%–74% Below Passing, 75%–82% Good, 83%–89% Very Good and 90%–100% Excellent. In line with institutional standards, the passing benchmark was 75%, equivalent to a raw score of 23 out of 30. Effect size was computed as Cohen's d for paired samples (the mean gain divided by the standard deviation of the gains) and interpreted as small (0.2), medium (0.5) or large (0.8) [13], to show the practical significance of the change in addition to its statistical significance [14].

The reflection journals were analyzed through inductive thematic analysis following the six phases of Braun and Clarke (2006): familiarization, initial coding, searching for themes, reviewing themes, defining and naming themes, and producing the report [15]. Qualitative data analysis is an interactive and interpretive process of data immersion, coding and theme development [16]. Each journal entry was read repeatedly and coded manually, similar codes were grouped into themes, and the themes were checked against the original entries. Trustworthiness was supported through member checking, audit trails linking codes to participant codes, frequency counts across the 19 participants, peer debriefing within the research team, and triangulation with the quantitative results.

2.5 Procedure for Implementation and Monitoring

The intervention was implemented over four weeks, each focused on one phase of the CREATE Approach (Figure 3). In Week 1 (Critically Reflect & Explore), students took the pretest, took mini-test simulations (Tests A, B and C), discussed fairness, clarity and accuracy, answered Discovery Guide Questions and created Concept Summary Charts. In Week 2 (Examine & Analyze), students examined sample tests and rubrics, recorded validity and reliability errors in an Error Analysis Chart and co-created a Framework for Good Test Design. In Week 3 (Apply & Test), groups designed mini-tests or performance rubrics in the Assessment Design Challenge and exchanged them for peer testing and validation using a Peer Evaluation Checklist. In Week 4 (Evaluate & Exhibit), students revised their tools based on peer and teacher feedback, took the posttest, wrote reflection journals and presented their outputs in the CREATE Exhibit, which was evaluated with a Verification Rubric. Outputs in every phase were scored with rubrics to ensure consistent scoring.

2.6 Ethical Considerations and Risk Management

Participants were fully informed of the purpose and flow of the study, participated voluntarily and gave their consent before any data were gathered. Participants were identified only by codes (S1 to S19), and individual results were kept confidential. Tests, journals and worksheets were stored securely, with digital files on a password-protected device and printed documents in a locked cabinet; the data will be kept for five years and then permanently destroyed. A risk assessment identified potential challenges in advance. Most risks, such as limited class time, difficulty engaging in peer testing, insufficient copies of materials, ambiguity in scenario-based items and inconsistent scoring during peer evaluation, were rated low and addressed through preventive actions such as expert review of materials, training students on rubric criteria and teacher monitoring of scoring. Class suspension due to weather or unforeseen events and power outages were rated high (risk score 12) and were addressed through make-up sessions, offline copies of materials and manual alternatives to digital activities.

3. The CREATE Intervention

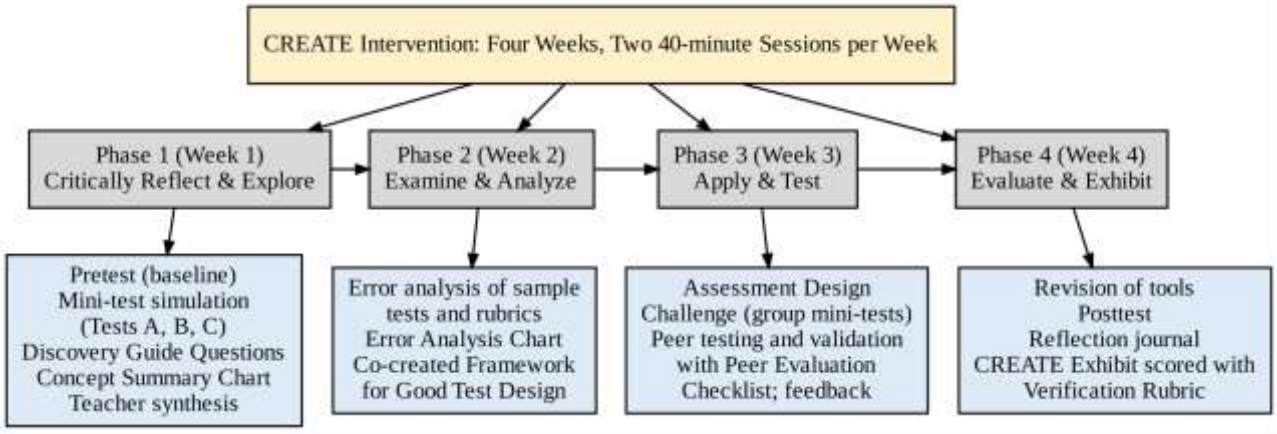
The CREATE Approach is a structured instructional intervention designed to address students' difficulties in analyzing situational questions, differentiating validity from reliability and applying both concepts to real assessment examples under limited time. Its name reflects the learning cycle it follows: Critically

Reflect, Explore, Analyze, Test and Evaluate. These elements are delivered in four phases, each targeting one root cause.

3.1 Components of the CREATE Approach

1. **Critically Reflect & Explore (Week 1; RC1).** Students took a scenario-based pretest and then completed contextualized tasks related to their major. Through the Bullseye Analogy, in which hitting the center illustrates validity and hitting the same spot repeatedly illustrates reliability, students moved from memorizing definitions to making sense of the concepts. Mini-test simulations, Discovery Guide Questions and Concept Summary Charts documented their initial understanding for later comparison.
2. **Examine & Analyze (Week 2; RC2).** In two collaborative sessions, students examined problematic assessment tools, such as poorly aligned rubrics and unclear test items, identified validity and reliability errors in an Error Analysis Chart, suggested corrections and co-created a Framework for Good Test Design.
3. **Apply & Test (Week 3; RC3).** Students used the framework to design 10-item mini-tests or rubrics in groups, justified their validity and reliability, and exchanged them with other groups for peer testing and validation. Feedback was documented, scoring consistency was checked, and outputs were revised.
4. **Evaluate & Exhibit (Week 4; RC4).** Students took the posttest under controlled time, wrote reflection journals on their conceptual development and presented their revised tools and reflection portfolios in the CREATE Exhibit, where groups explained how validity and reliability were demonstrated and the teacher evaluated outputs with the Verification Rubric.

Figure 3: Process Map of the CREATE Intervention



3.2 Resources, Quality Assurance and Sustainability

The researchers served as facilitators, material developers and data collectors, with planning support from the course instructor and validation support from subject-matter experts; the students were participants and peer evaluators. Materials included contextualized guides, task cards, Discovery Guide Questions, Error Analysis Charts, assessment design templates, Peer Evaluation Checklists, Verification Rubrics, tracking forms and sample tests. The itemized instructional materials cost Php 1,240 and were funded by the researchers, with a contingency fund of Php 500 for unforeseen expenses. Sessions were held in the Dressmaking Laboratory, arranged for group work and the exhibition of outputs. Implementation fidelity was monitored through protocol checks, time and task completion records and review of outputs, and delivery was refined weekly based on performance data, reflection journals and the researchers'

observation journal. Success indicators were completion of at least 90% of required outputs, 95% adherence to peer evaluation protocols, 90% peer scoring consistency by the third week, higher posttest than pretest scores for all students, and correct identification of validity and reliability issues by at least 90% of students in the error analysis. For sustainability, a CREATE Implementation Guide containing the rubrics, templates and test questionnaires was prepared, and the approach can be adapted to other assessment topics such as objective test construction and rubric development.

4. Results and Discussion

4.1 Baseline Knowledge of Validity and Reliability

Table 4 presents the individual pretest and posttest scores. Before the intervention, pretest scores ranged from 8 to 22 out of 30, with a mean of 16.79 (SD = 3.87). Five students (26.32%) were at the Very Poor level, 10 (52.63%) at the Poor level and four (21.05%) at the Below Passing level, so none of the 19 students reached the 75% passing standard (Table 5). The wide spread of scores shows that students entered the intervention with uneven levels of preparation.

Table 4: Pretest and Posttest Scores of the Students (n = 19; total score = 30)

Student	Pretest	%	Level	Remark	Posttest	%	Level	Remark	Gain
S1	18	60.00%	Poor	Failed	24	80.00%	Good	Passed	6
S2	21	70.00%	Below Passing	Failed	25	83.33%	Very Good	Passed	4
S3	12	40.00%	Very Poor	Failed	23	76.67%	Good	Passed	11
S4	18	60.00%	Poor	Failed	28	93.33%	Excellent	Passed	10
S5	14	46.67%	Very Poor	Failed	25	83.33%	Very Good	Passed	11
S6	18	60.00%	Poor	Failed	26	86.67%	Very Good	Passed	8
S7	8	26.67%	Very Poor	Failed	26	86.67%	Very Good	Passed	18
S8	16	53.33%	Poor	Failed	25	83.33%	Very Good	Passed	9
S9	22	73.33%	Below Passing	Failed	25	83.33%	Very Good	Passed	3
S10	14	46.67%	Very Poor	Failed	23	76.67%	Good	Passed	9
S11	16	53.33%	Poor	Failed	26	86.67%	Very Good	Passed	10
S12	19	63.33%	Poor	Failed	23	76.67%	Good	Passed	4
S13	17	56.67%	Poor	Failed	25	83.33%	Very Good	Passed	8
S14	21	70.00%	Below Passing	Failed	28	93.33%	Excellent	Passed	7
S15	22	73.33%	Below Passing	Failed	26	86.67%	Very Good	Passed	4
S16	10	33.33%	Very Poor	Failed	27	90.00%	Excellent	Passed	17
S17	16	53.33%	Poor	Failed	25	83.33%	Very Good	Passed	9

Student	Pretest	%	Level	Remark	Posttest	%	Level	Remark	Gain
S18	19	63.33%	Poor	Failed	28	93.33%	Excellent	Passed	9
S19	18	60.00%	Poor	Failed	24	80.00%	Good	Passed	6

The thematic analysis of the initial reflection notes supported this result. Students showed conceptual awareness of why validity and reliability matter, expressed in six themes: ensuring fairness in assessment, building trust and confidence, supporting teaching and learning, accuracy and meaningful measurement, the interconnection of validity and reliability, and potential confusion due to poor test design (Table 8). For example, one student wrote, "Tests help educators evaluate the effectiveness of their teaching strategies and adjust them as needed," and another, "Validity and reliability must work hand in hand. A test must consistently measure the right thing." However, this awareness had not yet translated into the ability to apply the concepts, which explains why all students failed the pretest.

4.2 Challenges in Analyzing Situational Questions

Students' main challenges, identified in the survey before the intervention, were analyzing situational questions (15 of 19 students, 78.95%), distinguishing validity from reliability (5 students, 26.32%) and time pressure when answering situational questions (3 students, 15.79%). These challenges persisted at the start of the intervention: in the post-intervention reflection journals, all 19 students described initial difficulty, as in "I encountered difficulties mainly because validity and reliability are closely related and sometimes appear similar." During the Examine & Analyze phase, students worked through these challenges by locating errors in sample tests and rubrics. All 19 students scored 17 to 19 out of 20 in the Error Analysis Chart (mean = 17.26, SD = 0.65), which corresponds to 85% to 95% and the Excellent level of the rubric (Table 6). Sixteen students scored 17, one scored 18 and two scored 19, which shows that although students met the expectations of the task, some aspects of reasoning and interpretation still needed clarification and reinforcement.

4.3 Contribution of Peer Review

In the Peer Testing Validation task, students evaluated the assessment tools of other groups for fairness, alignment, clarity and consistency and gave constructive feedback. All 19 students scored 17 or 18 out of 20 (mean = 17.63, SD = 0.50), corresponding to 85% to 90% and the Excellent level of the rubric (Table 6); 12 students scored 18 and seven scored 17. The narrow range of scores suggests that students shared an understanding of the criteria and were able to judge their peers' work consistently. The reflection journals support this result: 15 students described peer support as helpful, for example, "Discussing and validating answers with my peers helped me understand the criteria better." Peer review therefore strengthened the intervention by giving students repeated practice in applying the criteria and by providing feedback that they used to revise their outputs.

Table 5: Distribution of Students by Performance Level in the Pretest and Posttest

Performance Level	Percentage Range	Pretest f	Pretest %	Posttest f	Posttest %
Excellent	90%–100%	0	0.00	4	21.05
Very Good	83%–89%	0	0.00	10	52.63
Good	75%–82%	0	0.00	5	26.32

Performance Level	Percentage Range	Pretest f	Pretest %	Posttest f	Posttest %
Below Passing	70%–74%	4	21.05	0	0.00
Poor	50%–69%	10	52.63	0	0.00
Very Poor	0%–49%	5	26.32	0	0.00
Total		**19**	**100.00**	**19**	**100.00**

Table 6: Descriptive Statistics of the Error Analysis and Peer Testing Validation Scores (n = 19; total score = 20)

Statistic	Error Analysis Chart	Peer Testing Validation
Mean	17.26	17.63
Standard deviation	0.65	0.50
Median	17	18
Mode	17	18
Minimum–maximum	17–19	17–18
Percentage range	85%–95%	85%–90%
Rubric level (17–20 points)	Excellent (19 of 19)	Excellent (19 of 19)

4.4 Improvement in Analyzing Situational Questions

After the intervention, posttest scores ranged from 23 to 28, with a mean of 25.37 (SD = 1.61). All 19 students (100%) reached the passing standard: five (26.32%) were at the Good level, 10 (52.63%) at the Very Good level and four (21.05%) at the Excellent level (Tables 4 and 5). Every student improved, with gains ranging from 3 to 18 points (mean gain = 8.58, SD = 3.99). The lower standard deviation in the posttest shows that performance became both higher and more consistent across students. The largest gains were made by students with the lowest pretest scores (S7, +18; S16, +17), while the smallest gains were made by students who started near the passing standard (S9, +3; S2, S12 and S15, +4).

The paired-sample t-test (Table 7) showed that the difference between the pretest and posttest scores was statistically significant, $t(18) = 9.37$, $p < 0.001$ (two-tailed). The null hypothesis was therefore rejected. The 95% confidence interval of the mean gain was 6.66 to 10.50 points, and Cohen's d for paired samples was 2.15, which is a large effect.

Table 7: Paired-Sample t-Test of Pretest and Posttest Scores

Test	n	Mean	SD	Mean Gain	t	df	p (two-tailed)	Cohen's d	Decision
Pretest	19	16.79	3.87	8.58	9.37	18	< 0.001	2.15 (large)	Reject H0
Posttest	19	25.37	1.61						

The post-intervention reflection journals explain these gains. Six themes emerged (Table 8). Students described initial difficulty in distinguishing the two closely related concepts (19 students), the CREATE Approach as a structured guide (18 students), for example, "The CREATE approach has helped me better understand the difference between validity and reliability by giving me a step-by-step way to analyze information," development of critical thinking (17 students), for example, "When answering situational questions, the CREATE approach guided me to think critically instead of guessing," growth in confidence in evaluating tests and justifying answers (16 students), peer support (15 students), and coping strategies such as practice, repetition and reflection (12 students).

Table 8: Themes from the Students' Initial and Post-Intervention Reflections

Stage	Theme	Representative Excerpt	f
Initial	Ensuring fairness in assessment	"Reliable questions make the test fair…"	6
Initial	Building trust and confidence	"Credible assessment lends credibility to the institution…"	6
Initial	Supporting teaching and learning	"Tests help educators evaluate the effectiveness of their teaching strategies…"	8
Initial	Accuracy and meaningful measurement	"A valid test gives accurate results, helps teachers know if students really learned the lesson…"	10
Initial	Interconnection of validity and reliability	"Validity and reliability must work hand in hand."	6
Initial	Potential confusion due to poor test design	"Poorly constructed tests can lead to confusion, unfair assessment, and demotivation."	4
Post	Initial difficulty	"…validity and reliability are closely related and sometimes appear similar."	19
Post	CREATE as a structured guide	"…giving me a step-by-step way to analyze information."	18
Post	Critical thinking development	"…guided me to think critically instead of guessing."	17
Post	Confidence growth	"I feel more confident evaluating tests and justifying my answers now."	16
Post	Peer support	"Discussing and validating answers with my peers helped me understand the criteria better."	15
Post	Coping strategies	"Practice, repetition, and reflection helped me overcome my initial mistakes."	12

Note: f is the number of students whose reflections contained the representative code; a student's reflection may be counted under more than one theme.

4.5 Integration of Quantitative and Qualitative Findings

The quantitative data show that students' initial understanding of validity and reliability was low and uneven, with no student reaching the passing standard. During the intervention, all students reached the Excellent level in the error analysis and peer testing tasks with little variation, and the posttest showed higher scores, lower variability, a statistically significant gain and a large effect size. The qualitative findings are consistent with these trends. The initial reflections showed that students recognized the importance of fairness, accuracy and consistency in assessment but could not yet clearly distinguish and apply the concepts. After the intervention, their reflections described clearer distinctions, more organized and critical thinking, greater confidence in justifying answers and the value of practice, peer interaction and guided reflection. Together, the two strands indicate that the CREATE Approach addressed both the conceptual gaps and the practical difficulties of the students.

4.6 Discussion

Students commonly struggle to apply validity and reliability when designing classroom assessments, which reflects a gap between what they learn in theory and how they apply it in practice. Understanding validity and reliability is essential to ensure that assessment results accurately measure student learning and support instructional decisions [17], yet pre-service teachers often have limited assessment literacy, which makes it difficult to apply these concepts in real assessment tasks and calls for structured and practical learning interventions [3]. The CREATE Approach addressed this gap with a structured, hands-on framework that guided students through reflection, analysis, application and evaluation and allowed them to connect assessment concepts to real classroom tasks. This is consistent with findings that practicum experiences help future teachers apply theory to practice and develop their assessment literacy [18], and that assessment courses which include reflection and guided activities help pre-service teachers construct valid and reliable assessments [19]. By integrating interactive practice, collaborative activities and authentic assessment tasks, the approach helped students build the skills and confidence to create fair and consistent assessments.

These findings should be interpreted with caution. The study used one small, purposively selected group without a comparison group, so the improvement cannot be attributed to the CREATE Approach alone; regular coursework, maturation and the selection of low-scoring students may also have contributed. The same 30-item test was used as the pretest and the posttest about four weeks apart, so familiarity with the items may have raised posttest scores. In addition, the researchers facilitated the sessions and scored the outputs, and the qualitative data came from reflection journals written after a guided intervention, which may favor positive accounts.

5. Conclusion

The CREATE Approach (Critically Reflect, Explore, Analyze, Test, and Evaluate) effectively improved second-year BTLEd-HE students' understanding and application of validity and reliability in assessment. Students initially showed low performance and confusion in applying assessment principles, but after the intervention all of them reached the passing standard, most at the Very Good or Excellent level, and the gain was statistically significant with a large effect size. Reflections showed clearer understanding, more structured thinking and greater confidence. Grounded in experiential, constructivist, social, cognitive load and formative assessment theories, the CREATE Approach helped bridge theoretical knowledge and practical application and strengthened students' assessment competence.

5.1 Theoretical Implication

The study shows that structured, experiential and reflective interventions can help bridge the gap between theory and practice in assessment literacy. The CREATE Approach applies the experiential learning cycle of concrete experience, reflection, conceptualization and active experimentation; Constructivism through learners' active participation in building knowledge through engagement, cooperation and reflection; Formative Assessment Theory through continuous feedback and revision cycles; Social Learning Theory through peer interaction and collaborative validation; and Cognitive Load Theory through the breaking down of complex evaluation tasks into smaller, sequenced steps. Integrating these frameworks into instructional design can improve pre-service teachers' ability to apply validity and reliability in real classroom situations.

5.2 Sustainability and Action Plan

To sustain the improvement and prevent the recurrence of students' difficulties, monitoring will continue. Students who need additional guidance, including latecomers and those who need more examples, will receive individual support through small-group discussions and consultations. To accommodate short attention spans, activities will be organized into shorter, clearly focused segments with context-based examples. Mini-assessments, reflection journals and peer evaluation reports will be used to monitor progress, and a tracking system will check that students continue to apply assessment concepts in later courses. For future batches, pre-prepared instructional materials, multiple scenario-based exercises, rubrics, mini-tests and a structured timeline integrating practice, collaboration and reflection will be maintained, and flexible review opportunities will allow latecomers to follow the intervention without missing important material. Lessons will remain modular and experiential, with each CREATE stage implemented as a self-contained unit targeting specific learning objectives.

5.3 Reflection and Recommendations

The intervention showed that structured, experiential activities can help students grasp abstract concepts such as validity and reliability, and that multiple examples and peer collaboration support students who learn at different paces. Short, focused tasks combined with reflection helped sustain attention and deepen learning. It is therefore recommended that teachers continue to use systematic experiential activities, scenario-based tasks, collaborative peer evaluation and reflection journals to develop assessment literacy; provide individual guidance, varied illustrative examples and supplementary materials, including digital materials, for latecomers and students who need more practice; and use flexible strategies to maintain the engagement of students with short attention spans. Future studies should use a comparison group, a parallel form of the test for the posttest and a larger number of participants, and should examine whether the gains are retained in later courses and teaching practice.

Acknowledgement

The researchers express their sincere gratitude to Dr. Maria Gloria R. Lugo, research adviser, for her unwavering support, guidance and valuable feedback; to the members of the research panel, Cheramie M. Ambayec, James Roland V. Galvez, Guillermo L. Bello Jr. and Cherilyn V. Malintad; to Eddie Jr. S. Panggo, critique teacher of the BTLEd faculty, for his constructive feedback and recommendations; to the subject-matter experts who validated the instruments; and to their classmates for their active participation in the implementation of the CREATE Approach.

References

1. Commission on Higher Education, "Policies, Standards and Guidelines for the Bachelor of Technology and Livelihood Education (BTLEd)", CHED Memorandum Order No. 78, Series of 2017. https://ched.gov.ph/wp-content/uploads/2017/11/CMO-No.-78-s.2017.pdf
2. Department of Education, "National Adoption and Implementation of the Philippine Professional Standards for Teachers", DepEd Order No. 42, Series of 2017, August 2017. https://www.deped.gov.ph/2017/08/11/do-42-s-2017-national-adoption-and-implementation-of-the-philippine-professional-standards-for-teachers/
3. S. Pastore, "Teacher Assessment Literacy: A Systematic Review", Frontiers in Education, July 2023, 8, Article 1217167. https://doi.org/10.3389/feduc.2023.1217167
4. W.O.S. Wardani, D. Sukyadi, P. Purnawarman, "Exploring Teacher Assessment Literacy in EFL Classroom", Proceedings of the 2nd International Conference on Progressive Education (ICOPE 2020), European Alliance for Innovation, 2021. https://doi.org/10.4108/eai.16-10-2020.2305253
5. A. Arhin, A. Yeboah, "Teachers' Conceptualization of Validity and Reliability in Classroom Assessment in Ghana", African Journal of Teacher Education, June 2025, vol. 14, no. 1, June 2025, pp. 87–119. https://doi.org/10.21083/ajote.v14i1.8038
6. S. McLeod, "Constructivism as a Theory for Teaching and Learning", Simply Psychology, March 2025. https://www.simplypsychology.org/constructivism.html
7. Institute for Experiential Learning, "What Is Experiential Learning?", 2025. https://experientiallearninginstitute.org/what-is-experiential-learning/
8. S. McLeod, "Albert Bandura's Social Learning Theory", Simply Psychology, October 2025. https://www.simplypsychology.org/bandura.html
9. V. Likourezos, "An Introduction to Cognitive Load Theory", The Education Hub, March 2021. https://theeducationhub.org.nz/an-introduction-to-cognitive-load-theory/
10. P. Black, D. Wiliam, "Developing the Theory of Formative Assessment", Educational Assessment, Evaluation and Accountability, 2009, 21 (1), 5–31. https://doi.org/10.1007/s11092-008-9068-5
11. T. George, "Mixed Methods Research: Definition, Guide and Examples", Scribbr, August 2021. https://www.scribbr.com/methodology/mixed-methods-research/
12. Munawaroh, N.S. Setyani, L. Susilowati, "The Development of Electronic Problem Based Learning (E-PBL) Method in Entrepreneurship Course", Atlantis Press, November 2022. https://doi.org/10.2991/978-2-494069-15-2_15
13. J. Cohen, Statistical Power Analysis for the Behavioral Sciences, 2nd Edition, Lawrence Erlbaum Associates, 1988.
14. Statsig Team, "5 Things to Understand about Statistical Significance", Statsig, July 2024. https://www.statsig.com/perspectives/things-to-understand-about-statistical-significance
15. V. Braun, V. Clarke, "Using Thematic Analysis in Psychology", Qualitative Research in Psychology, 2006, 3 (2), 77–101. https://doi.org/10.1191/1478088706qp063oa
16. N. Dahal, "Qualitative Data Analysis: Reflections, Procedures, and Some Points for Consideration", Frontiers in Research Metrics and Analytics, September 2025. https://doi.org/10.3389/frma.2025.1669578
17. S. Kime, "Four Pillars of Assessment: Validity", Evidence Based Education, November 2017. https://evidencebased.education/resource/pillars-assessment-purpose-validity/

18. R. Pardo, D. García-Pérez, E. Panadero, "Shaping the Assessors of Tomorrow: How Practicum Experiences Develop Assessment Literacy in Secondary Education Pre-Service Teachers", Teaching and Teacher Education, 2024, 152, Article 104798. https://doi.org/10.1016/j.tate.2024.104798
19. H.K.G. Alcazaren, "Developing Language Assessment Literacy of Filipino Pre-Service Teachers: Reflections from a Language Assessment Course", TEFLIN Journal, December 2024, 35 (2), 216–231. https://doi.org/10.15639/teflinjournal.v35i2/216-231