Asian Journal of Basic Science & Research
Volume 7, Issue 3, Pages 226-243, July-September 2025

Development and Content Validation of a Critical Thinking Scale for B.Ed. Students

M.C. Yarriswamy[1] & Archana Pujar[2*]

[1]*Professor, Department of Education, Bangalore University, Bengaluru, Karnataka, India.* [2]*Research Scholar, Department of Education, Rani Channamma University, Belagavi, Karnataka, India. Email: pujar.archana.h@gmail.com*[*]

DOI: Under Assignment

Article Received: 19 July 2025 *Article Accepted: 25 September 2025* *Article Published: 28 September 2025*

ABSTRACT

Purpose: This study aimed to develop and validate a Critical Thinking Scale (CTS) for Bachelor of Education (B.Ed.) students by integrating expert-based content validation with pilot psychometric testing.

Design/Approach/Methods: An initial pool of 34 items was generated from critical thinking theory, teacher education literature, and policy frameworks (NEP 2020). An expert panel of 11 members evaluated items using the Content Validity Index (CVI) and Content Validity Ratio (CVR). Items meeting both thresholds (I-CVI ≥ 0.78; CVR ≥ 0.59) were retained. The refined scale was pilot-tested with 89 B.Ed. students, and data were analysed for reliability, item–total correlations, and discrimination indices. Raven's Advanced Progressive Matrices (RAPM) were administered to confirm sampling equivalence.

Findings: Expert review reduced the pool of items, with the scale-level CVI improving to 0.874. Pilot testing confirmed this refinement: Cronbach's α improved from 0.78 (34 items) to 0.80 (25 items). Item analysis identified weak items with low discrimination or ceiling effects, aligning with expert recommendations. RAPM scores were symmetrically distributed, supporting matched-pair sampling for future experimental designs.

Originality/Value: The study introduces a contextually relevant, psychometrically sound CTS for Indian teacher education, addressing both theoretical and policy needs. The validated item scale provides a foundation for large-scale validation and cross-institutional application.

Keywords: Critical Thinking; Teacher Education; Scale Development; Instrument Validation; Content Validity; Content Validity Index; Content Validity Ratio; Item Analysis; Reliability Testing; Psychometric Validation.

1. Introduction

Critical thinking (CT) is a cornerstone of higher education, yet within teacher education it carries particular significance. Pre-service teachers must not only acquire disciplinary knowledge but also demonstrate the ability to interpret classroom contexts, evaluate evidence, and make pedagogical decisions that promote equitable and effective learning (Ennis, 2011; Dwyer, Hogan, & Stewart, 2014). The Delphi Report defined CT as "purposeful, self-regulatory judgment" encompassing interpretation, analysis, evaluation, inference, and explanation (Facione, 1990). Complementing these skill dimensions, dispositions such as open-mindedness, reflectiveness, and truth-seeking remain central to teachers' professional identity and decision-making processes (Tiruneh et al., 2017). Together, these perspectives underscore CT's multidimensional nature and its status as an indispensable attribute for teacher candidates.

Yet, despite its prominence in policy frameworks and curricular documents including India's National Education Policy (NEP) 2020, which explicitly emphasizes higher-order thinking - the valid measurement of CT among pre-service teachers remains a persistent challenge. Existing instruments such as the Watson–Glaser Critical Thinking Appraisal and the California Critical Thinking Skills Test are widely used in Western contexts but are costly, language-dependent, and culturally bounded, thus limiting their applicability in Indian teacher education settings (Yuan et al., 2023). Locally developed scales, while more context-sensitive, often lack rigorous validation, raising concerns about construct coverage and psychometric adequacy. As a result, teacher education institutions

are left without reliable tools to assess either baseline CT competencies or the impact of instructional interventions aimed at fostering them.

Establishing content validity is the crucial first step in addressing this measurement gap. Among the various approaches, two indices have achieved widespread recognition: The Content Validity Index (CVI), proposed by Lynn (1986), which quantifies the degree of expert consensus on item relevance; and the Content Validity Ratio (CVR), developed by Lawshe (1975), which assesses the essentiality of items based on expert judgments. CVI ensures that items are perceived as relevant representations of the construct, while CVR ensures that items are indispensable to the construct domain. The combination of CVR and CVI thus provides a dual-filter mechanism that balances comprehensiveness with parsimony (Polit & Beck, 2006). In the present work, this combined procedure proved particularly effective: the initial Critical Thinking pool of items was refined, with an increase in mean I-CVI from 0.856 to 0.874, reflecting higher expert agreement on retained items.

However, content validation alone is insufficient. Items that appear valid to experts may fail to demonstrate psychometric robustness when administered to the target population (Messick, 1994). Empirical validation requires a sequence of procedures including pilot administration, item analysis, reliability testing, and factor analytic modelling (Worthington & Whittaker, 2006; Brown, 2015). Findings from our pilot study highlight this imperative: while the Communication and Collaboration subscales exhibited acceptable reliability ($\alpha \geq 0.70$), the Critical Thinking subscale yielded an α of 0.48, indicating poor internal consistency and underscoring the necessity of iterative refinement.

Against this backdrop, the present study integrates content validation (CVI and CVR) with instrument validation techniques to produce a robust and contextually appropriate Critical Thinking Scale for B.Ed. students. Specifically, the objectives are: (a) to refine CT items through expert consensus using combined CVR–CVI thresholds, and (b) to establish a psychometric validation roadmap that ensures reliability and construct validity in subsequent large-scale applications. This dual contribution not only addresses methodological gaps in instrument development but also provides teacher education institutions with a validated tool aligned with policy imperatives and international best practices.

1.1. Study Objectives

The study was guided by the following objectives:

1) To develop an initial item pool for assessing critical thinking in B.Ed. students based on theoretical and curricular frameworks.

2) To establish content validity of the CTS using CVR and CVI indices with expert review.

3) To conduct a pilot study for item analysis and reliability estimation of the CTS.

4) To refine the scale by removing or revising weak items, ensuring improved psychometric strength.

5) To test the feasibility of matched-pair sampling using Raven's Advanced Progressive Matrices (RAPM).

6) To provide a validated items CTS as a foundation for future large-scale validation studies.

2. Literature Review

2.1. Conceptualizing Critical Thinking in Teacher Education

Critical thinking is widely conceptualized as a multi-faceted construct that combines cognitive skills (interpretation, analysis, inference, evaluation, explanation) with dispositional attributes (open-mindedness, intellectual humility, reflective scepticism) (Facione, 1990). Modern treatments emphasize CT as both a set of teachable skills and a professional disposition that enables teachers to interpret classroom evidence, design adaptive instruction, and evaluate learning trajectories (Ennis, 2011; Dwyer, Hogan, & Stewart, 2014). Recent bibliometric and review work confirms CT's centrality in teacher education research and highlights trends toward contextually sensitive conceptualizations that couple cognitive processes with pedagogical practices.

2.2. Measurement Approaches and Their Limitations

Historically, CT has been measured via two broad approaches: (a) standardized, performance-based tests (e.g., Watson–Glaser, CCTST) intended to assess cognitive skill proficiency, and (b) self-report or disposition inventories that gauge attitudes and tendencies toward reflective reasoning (Facione, 1990; Payan-Carreira et al., 2022). While performance tests offer criterion-referenced tasks, they often suffer from cultural and linguistic bias, high licensing costs, and limited alignment with local curricular practices; conversely, self-report measures can inflate perceived ability and fail to capture actual reasoning performance (Payan-Carreira et al., 2022). Galindo-Domínguez et al. (2023) validated a 42-item CT scale grounded in teachers' perspectives, organized into six dimensions, and demonstrated strong psychometric properties across diverse university cohorts. Similarly, Rodríguez-Rojas (2024) developed the Critical Thinking Evaluation Scale (CTES) in Colombia, offering a culturally sensitive instrument with robust validity evidence. These limits are especially acute in non-Western, resource-constrained teacher education contexts—motivating the need for locally developed, validated instruments.

2.3. Content Validation: CVI and CVR - Theory and Recent Advances

Content validity is the foundational evidentiary strand in scale development: it demonstrates that the item set adequately samples the intended construct domain (Lawshe, 1975; Lynn, 1986). Two widely used indices are the Item-level Content Validity Index (I-CVI) and Scale-level CVI (S-CVI) (Lynn, 1986), which quantify expert agreement on item relevance, and Lawshe's Content Validity Ratio (CVR), which measures essentiality (Lawshe, 1975). Contemporary methodological reviews recommend reporting I-CVI, S-CVI/Ave, and adjusted kappa values to account for chance agreement, and they caution against sole reliance on S-CVI/UA (Polit & Beck; Ayre & Scally). Recent work revisiting Lawshe's method has defended its utility while recommending transparent reporting of critical values and panel composition.

2.4. Best Practices in Applying CVI and CVR

Recent methodological guidance and empirical studies converge on several best practices: (a) assemble a heterogeneous expert panel (content experts, assessment methodologists, and practitioners) of adequate size (often 8-15) to balance diversity and decision-rule stability; (b) use a multi-dimensional rating rubric (e.g., relevance +

essentiality) and compute both I-CVI and CVR; (c) report both S-CVI/Ave and chance-corrected indices (modified kappa); and (d) document explicit decision rules (retain if both I-CVI and CVR thresholds are met) and reasons for item removal or revision (Polit & Beck, 2006; Romero Jeldres, 2023; Zamanzadeh et al., 2015). These practices reduce ambiguity and improve reproducibility of content validation outcomes.

2.5. From Content Validity to Psychometric Validation: Classical and Modern Workflows

Content validation is a necessary but not sufficient step. Scale developers are advised to proceed through a staged empirical pipeline: pilot administration to target respondents → classical item analysis (difficulty, discrimination, item-total correlations) → internal consistency estimation (Cronbach's α and McDonald's ω) → exploratory factor analysis (EFA; to identify latent structure) → confirmatory factor analysis (CFA; to test and cross-validate structure) → criterion-related and convergent/discriminant evidence (Boateng et al., 2018; Worthington & Whittaker, 2006; Brown, 2015). Recent primers emphasize parallel analysis, oblique rotations for correlated constructs, use of omega and item response theory (IRT) where appropriate, and adequate sample sizes (e.g., ≥200 for EFA; larger for CFA/SEM depending on model complexity).

2.6. Contemporary Studies Validating Critical Thinking Measures

Several recent studies demonstrate practical routes to short, psychometrically sound CT measures. For example, Payan-Carreira et al. (2022) developed and validated a short form of a comprehensive CT self-assessment scale using combined EFA/CFA procedures and reported acceptable reliability and factorial invariance across subgroups. Other domain-specific CT measures in STEM education and teacher training have used multi-stage validation processes (item generation → expert review → pilot testing → EFA/CFA) with favourable results, illustrating the feasibility of rigorous instrument construction in educational settings. These studies also highlight common pitfalls overreliance on small pilot Ns and failure to report chance-adjusted content indices that the current protocol aims to avoid.

Recent studies continue to refine CT measurement. For example, Galindo-Domínguez et al. (2023) validated a 42-item CT instrument based on six dimensions. Similarly, Rodríguez-Rojas (2024) introduced the CT Evaluation Scale (CTES) with strong validity in a Latin American context. Fabio et al. (2025) developed the CRA with excellent reliability (α = 0.93). Others, such as Hultquist et al. (2023) streamlined existing CTS via confirmatory factor analysis, while Önal (2025) tailored CT measurement specifically for pre-service teachers. These contemporary works reflect ongoing efforts to balance theoretical fidelity, empirical rigor, and contextual adaptability in CTS.

2.7. Critical Thinking in Teacher Education and the Policy Imperative

Policy documents, most notably India's National Education Policy (NEP) 2020, foreground higher-order thinking and critical reasoning as core educational outcomes for pre-service training and curricular reform. NEP 2020 underscores the need for teacher training programs to incorporate and assess critical thinking and problem-solving abilities (Government of India, 2020). Empirical assessments of NEP's impact on CT are nascent but indicate growing scholarly attention to aligning teacher education assessment with policy expectations (studies 2021-2024).

These policy drivers increase the demand for locally validated CT measurement tools that can inform both programme improvement and policy monitoring.

2.8. Gaps Identified in the Literature

Despite methodological advances, the literature shows recurring gaps relevant to the present study:

- Insufficient integration of content validation and empirical psychometrics. Many studies either report CVI/CVR without subsequent EFA/CFA, or they skip transparent content-validation reporting entirely (Romero Jeldres, 2023).
- Limited context-specific instruments for teacher education in low- and middle-income settings. Most widely used CT tests are developed in Western contexts and are not always suitable for Indian B.Ed. cohorts.
- Variable reporting of expert-panel methods and thresholds. Failure to report panel composition, decision rules, and chance-adjusted agreement statistics weakens reproducibility.
- Small pilot samples for factor analysis. Several validation studies rely on underpowered samples for EFA/CFA, undermining structural claims; best practice recommends larger pilot Ns and split-sample CFA.

These gaps collectively argue for a rigorous protocol that (a) combines CVR and CVI with explicit decision rules, (b) documents panel composition and chance-adjusted indices, and (c) follows through with ample pilot sampling, classical item analysis, and factor analytic validation, exactly the pipeline used in the present study.

Bringing together conceptual clarity on CT with methodological rigor in instrument development, the literature supports a dual-evidence approach: first, robust content validation using CVI and CVR (with chance-correction), and second, empirical psychometric testing via item analysis and factor modelling (Boateng et al., 2018; Romero Jeldres, 2023). The current research adopts this integrated stance and situates the resulting Critical Thinking Scale within policy needs (NEP 2020) and practical assessment requirements for B.Ed. programs. By explicitly combining expert-based essentiality (CVR) and relevance (I-CVI/S-CVI) criteria with standard psychometric pipelines (EFA/CFA, reliability, criterion validity), the study addresses long-standing methodological shortcomings and contributes a replicable model for future measurement work in teacher education.

3. Methods

3.1. Research Design

This study followed a sequential instrument development design (Boateng et al., 2018), which integrates both expert-based content validation and empirical psychometric validation. The process began with item generation informed by theoretical and policy frameworks, proceeded through structured expert review using Content Validity Index (CVI) and Content Validity Ratio (CVR) procedures, and was followed by pilot testing with the target population. The ultimate goal was to establish a reliable and valid Critical Thinking Scale for B.Ed. students.

3.2. Item Generation

An initial item pool was developed based on:

- The Delphi Report framework (Facione, 1990), which defines critical thinking as comprising skills such as interpretation, analysis, evaluation, inference, explanation, and self-regulation.
- Educational research in teacher education (Ennis, 2011; Dwyer et al., 2014), highlighting CT as central to professional classroom reasoning.
- Policy imperatives, especially India's National Education Policy (NEP, 2020), which emphasizes higher-order thinking in pre-service teacher education.

The preliminary item pool consisted of statements rated on a 5-point Likert scale (1 = Strongly Disagree to 5 = Strongly Agree), framed in accessible language for first-year B.Ed. students.

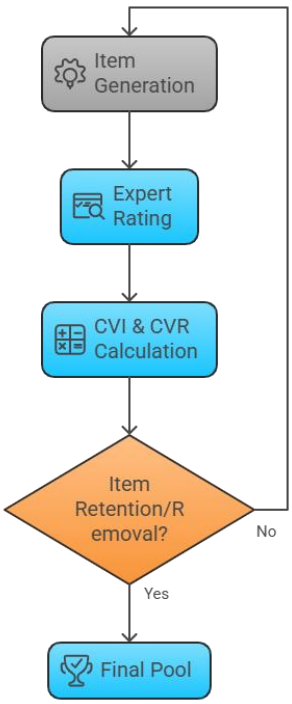

Figure 1. Content Validation Framework for the CTS (CVI and CVR procedures)

The flow diagram shown in figure 1 illustrates the content validation process used to refine the initial pool of Critical Thinking (CT) items. The process begins with item generation, drawing upon theoretical frameworks (Facione, 1990; Ennis, 2011), curricular standards, and empirical literature. Items are then subjected to expert panel review, where domain specialists rate each item's relevance and essentiality using a structured rubric. The ratings are quantified using Content Validity Index (CVI) and Content Validity Ratio (CVR). Items meeting both thresholds (I-CVI $\geq$ 0.78; CVR $\geq$ 0.59 for 11 experts) are retained, while others are either revised or removed. The process concludes with a refined item pool demonstrating improved consensus and parsimony.

3.3. Expert Panel

Content validity was assessed by a purposively selected panel of 11 experts, including:

- Teacher education faculty,
- Educational assessment specialists, and

- Experienced practitioners in teacher training.

Panellists had ≥10 years of experience in teaching, research, or psychometrics. This heterogeneity ensured balanced perspectives on both conceptual coverage and practical applicability of items.

3.4. Content Validation Procedure

Experts independently rated items on two dimensions:

- Relevance (CVI): Using a 4-point scale (1 = not relevant to 4 = highly relevant). Item-level CVI (I-CVI) and scale-level CVI (S-CVI/Ave) were computed (Lynn, 1986). A minimum I-CVI of 0.78 was required for retention.

- Essentiality (CVR): Using Lawshe's (1975) three-category scale (essential, useful but not essential, not essential). CVR was calculated using the standard formula, with critical values determined by expert panel size (e.g., 0.59 for 11 experts).

Decision rule: Items meeting both thresholds (I-CVI ≥ 0.78 and CVR ≥ critical value) were retained; others were flagged for revision or removal.

3.5. Pilot Testing

The refined scale was administered to a sample of B.Ed. students drawn from one teacher education institution. Data collection adhered to ethical guidelines, including informed consent and voluntary participation. The pilot was designed to enable item analysis (difficulty, discrimination, item–total correlation) and preliminary reliability assessment. Although sample size was limited, the pilot provided initial insights into instrument feasibility and internal consistency.

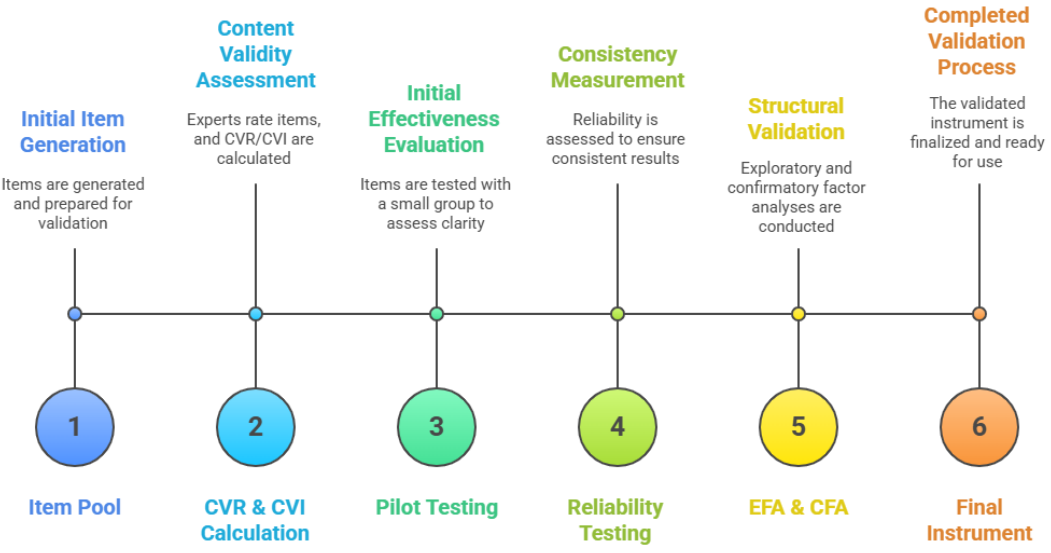

Figure 2. Pilot Testing and Item Refinement Workflow for the CTS

The figure 2 depicts the sequential steps of instrument validation following content refinement. The validated item pool undergoes pilot testing with the target population (B.Ed. students). Data from the pilot are analysed for item performance (difficulty, discrimination, item–total correlation) and internal consistency reliability (Cronbach's alpha, McDonald's omega). Items failing to meet psychometric criteria are flagged for revision or removal. The

next stage involves Exploratory Factor Analysis (EFA) to uncover underlying factor structures, followed by Confirmatory Factor Analysis (CFA) on independent samples to test model fit indices (e.g., CFI, RMSEA, SRMR). The workflow concludes with the establishment of a validated scale, ready for large-scale implementation and criterion-related validation (e.g., association with teaching performance or academic outcomes).

3.6. Planned Psychometric Validation

For large-scale validation, the following analyses are planned:

- Item analysis: Difficulty indices, discrimination indices, and corrected item–total correlations.
- Reliability estimation: Cronbach's α and McDonald's ω for internal consistency.
- Exploratory Factor Analysis (EFA): Using parallel analysis and oblique rotations to identify latent CT factors.
- Confirmatory Factor Analysis (CFA): To test factor structure and assess fit indices (CFI, RMSEA, SRMR).
- Convergent and discriminant validity: Correlations with related and unrelated constructs.
- Criterion validity: Associations between CT scores and academic or pedagogical performance indicators.
- Invariance testing: Where possible, examine measurement invariance across gender, language, and institutional groups.

This comprehensive approach ensures that the final Critical Thinking Scale meets contemporary standards of validity, reliability, and applicability.

3.7. Ethical Considerations

The study adhered to ethical standards for research involving human participants. Informed consent was obtained from all participants, who were assured of confidentiality, anonymity, and the voluntary nature of their involvement. Data were used solely for research purposes. Ethical clearance was obtained from the Institutional Research and Ethics Committee of the affiliated teacher education institution.

3.8. Participant Demographics

The pilot sample comprised 89 B.Ed. students enrolled in their first year of a two-year teacher education program at a government-aided college in India. The sample included 64% female and 36% male students, with an average age of 21.4 years (SD=1.9). Students represented diverse subject specializations (science, social science, and humanities), ensuring a balanced cohort reflective of typical teacher education enrolment.

4. Results

4.1. Content Validation (Expert Review)

The initial pool of 34 Critical Thinking items was evaluated by 11 experts using both CVI and CVR indices. Eleven experts were selected based on Lynn's (1986) guideline recommending 5-15 experts for CVI analysis, balancing feasibility with the need for stable consensus.

- Item-level CVI (I-CVI): Ranged between 0.70 and 0.91. Items scoring <0.78 were flagged.

- Content Validity Ratio (CVR): For 11 experts, the minimum acceptable CVR was 0.59. Items below this threshold were considered non-essential.
- After applying the combined CVR + CVI rule, the pool was reduced from 34 to 25 items.
- The scale-level CVI (S-CVI/Ave) improved from 0.856 to 0.874, reflecting stronger expert consensus.

Table 1. Content Validation Results for Critical Thinking Items

Stage	No. of Items	Mean I-CVI	Items ≥ I-CVI	Items ≥ CVR	Retained Items
Initial pool	34	0.856	28	27	34
After CVR + CVI review	25	0.874	25	25	25

Table 2. Item-wise Content Validation Results for Critical Thinking Scale (N = 11 Experts)

Item No.	I-CVI	CVR	Decision
CT1	0.91	0.64	Retain
CT2	0.82	0.55	Revise
CT3	0.88	0.73	Retain
CT4	0.79	0.61	Retain
CT5	0.72	0.55	Remove
CT6	0.85	0.67	Retain
CT7	0.80	0.59	Retain
CT8	0.83	0.62	Retain
CT9	0.76	0.58	Revise
CT10	0.87	0.64	Retain
CT11	0.90	0.70	Retain
CT12	0.78	0.61	Retain
CT13	0.81	0.65	Retain
CT14	0.75	0.55	Revise
CT15	0.88	0.71	Retain
CT16	0.80	0.59	Retain
CT17	0.82	0.60	Retain
CT18	0.70	0.54	Remove
CT19	0.84	0.63	Retain
CT20	0.86	0.66	Retain
CT21	0.77	0.57	Revise
CT22	0.83	0.62	Retain

CT23	0.85	0.68	Retain
CT24	0.80	0.59	Retain
CT25	0.74	0.55	Revise
CT26	0.89	0.72	Retain
CT27	0.78	0.60	Retain
CT28	0.71	0.54	Remove
CT29	0.82	0.63	Retain
CT30	0.84	0.65	Retain
CT31	0.73	0.55	Revise
CT32	0.87	0.70	Retain
CT33	0.90	0.73	Retain
CT34	0.85	0.67	Retain

4.2. Pilot Testing (Item Analysis & Reliability)

Following expert review, the original 34-item Critical Thinking Scale was administered to a pilot sample of 89 B.Ed. students to evaluate empirical item performance and internal consistency reliability. The pilot study served as the first test of the instrument's feasibility in a real educational context, and its results informed further item refinement.

4.2.1. Internal Consistency Reliability

The reliability of the instrument was estimated using Cronbach's alpha (α), which measures the internal consistency of items.

Table 3. Reliability Analysis of Critical Thinking Scale (Pilot Data)

Scale Version	No. of Items	Cronbach's α	Reliability Status
Original (all items)	34	0.78	Good
Refined (after analysis)	25	0.80	Good

The 34-item version yielded an α = 0.78, which meets the conventional threshold of 0.70 for acceptable reliability and approaches the 0.80 benchmark considered good in social science research (Nunnally & Bernstein, 1994). After item refinement, the 25-item version produced an α = 0.80, demonstrating that the shorter scale not only reduced redundancy but also improved internal consistency. This improvement suggests that removing weak or redundant items enhanced the coherence of the scale without compromising construct coverage.

4.2.2. Item Difficulty

Item difficulty was assessed by computing mean response values for each of the 34 items on a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree). Acceptable item means are generally expected to fall between 2.5 and 4.0, as values outside this range may indicate floor or ceiling effects.

Table 4. Descriptive Statistics of Item Difficulty for Critical Thinking Items (N = 89)

Range of Item Means	No. of Items	Item Nos.	Interpretation
3.60 – 3.79	10	CT2, CT7, CT11, CT13, CT16, CT19, CT22, CT23, CT27, CT31	Acceptable
3.80 – 3.99	14	CT3, CT4, CT6, CT10, CT12, CT15, CT17, CT20, CT24, CT26, CT29, CT30, CT32, CT34	Optimal range
4.00 – 4.19	5	CT8, CT9, CT21, CT25, CT33	Slightly high, retained after review
≥ 4.20	5	CT1, CT5, CT14, CT18, CT28	Ceiling effect → removed

In this study, item means ranged from 3.60 to 4.40, suggesting that most items were moderately to strongly endorsed by respondents. However, five items (CT1, CT5, CT14, CT18, CT28) had means above 4.20, indicating potential ceiling effects. Such items were endorsed by nearly all respondents, reducing their ability to differentiate between students with stronger and weaker critical thinking abilities.

Conversely, no item fell below the lower bound (<2.5), which indicates that items were not overly difficult or confusing for students to engage with. The distribution of means clustered around the midpoint (3.80–4.00), suggesting that the scale adequately captured variance in student responses.

Items identified with ceiling effects were subsequently removed or revised to improve scale discrimination. This contributed to the refinement of items, ensuring that the final scale retained items with optimal difficulty characteristics.

4.2.3. Item-Total Correlations

Item–total correlations assess how well each item aligns with the overall construct measured by the scale.

- Items with correlations ≥ 0.30 were considered acceptable contributors to the scale.
- Items falling below this threshold were marked for revision or removal.
- In the pilot, most retained items exhibited correlations between 0.34 and 0.46, showing adequate coherence with the overall construct of critical thinking.

4.2.4. Discrimination Indices

The discrimination index compares the performance of high-scoring and low-scoring groups (top 27% vs. bottom 27%).

- Items with D ≥ 0.30 were deemed effective discriminators.
- Items with lower discrimination values (<0.30) failed to distinguish between strong and weak critical thinkers and were eliminated.

4.2.5. Item Reduction Process

Based on the combined criteria of item difficulty, item–total correlations, and discrimination indices:

- Nine items were removed.
- The retained items demonstrated stronger alignment with the construct, balanced difficulty levels, and adequate discriminative capacity.

Table 5. Item Analysis for Critical Thinking Items

Item No.	Mean (Difficulty)	Item–Total Correlation	Discrimination Index	Decision
CT1	4.25	0.22	0.34	Remove
CT2	3.65	0.35	0.52	Retain
CT3	3.90	0.39	0.55	Retain
CT4	3.80	0.29	0.45	Revise
CT5	4.40	0.21	0.28	Remove
CT6	3.95	0.38	0.50	Retain
CT7	3.70	0.36	0.47	Retain
CT8	4.10	0.41	0.52	Retain
CT9	4.05	0.26	0.33	Revise
CT10	3.85	0.37	0.49	Retain
CT11	3.75	0.40	0.53	Retain
CT12	4.00	0.34	0.46	Retain
CT13	3.60	0.32	0.44	Retain
CT14	4.35	0.24	0.29	Remove
CT15	3.95	0.46	0.60	Retain
CT16	3.80	0.39	0.51	Retain
CT17	4.05	0.28	0.31	Revise
CT18	4.20	0.20	0.26	Remove
CT19	3.85	0.37	0.48	Retain
CT20	3.90	0.42	0.56	Retain
CT21	4.05	0.28	0.30	Remove
CT22	3.80	0.36	0.49	Retain
CT23	3.70	0.38	0.52	Retain
CT24	4.00	0.35	0.47	Retain
CT25	4.15	0.22	0.29	Remove
CT26	3.85	0.39	0.51	Retain
CT27	3.75	0.34	0.46	Retain
CT28	4.30	0.21	0.28	Remove
CT29	3.85	0.36	0.49	Retain
CT30	3.90	0.41	0.55	Retain
CT31	3.70	0.27	0.33	Revise
CT32	3.95	0.40	0.53	Retain
CT33	3.80	0.43	0.57	Retain
CT34	3.85	0.39	0.50	Retain

The pilot testing phase confirmed that the initial 34-item instrument had acceptable reliability (α = 0.78), but refinement through item analysis improved the psychometric quality. The final 25-item scale demonstrated good internal consistency (α = 0.80), appropriate item difficulty, and satisfactory discrimination indices. This convergence of expert judgment (CVR + CVI) and pilot psychometrics strengthens the validity of the scale and prepares it for further large-scale validation, including factor analysis and criterion testing.

5. Discussion

The present study set out to develop and validate a Critical Thinking (CT) Scale for B.Ed. students, integrating expert-based content validation with empirical pilot testing. The findings demonstrate that both stages of validation - expert consensus through CVR and CVI and student data through pilot analysis converged to refine the original item pool into a final set that exhibited stronger reliability and clearer construct coherence.

One of the most encouraging results was the internal consistency of the instrument. The original 34-item version already achieved an acceptable level of reliability (Cronbach's α = 0.78), suggesting that the pool was conceptually well constructed. However, reliability further improved to 0.80 after removing weak items, indicating that shorter, more targeted instruments can outperform longer scales when psychometric refinement is applied. This aligns with established psychometric principles, where redundancy or poorly discriminating items tend to suppress overall reliability (Nunnally & Bernstein, 1994). The outcome supports recent work on short-form Critical Thinking Scales that demonstrated strong psychometric performance while reducing respondent burden (Payan-Carreira et al., 2022).

The convergence of expert review and pilot results is particularly significant. During content validation, experts reduced the pool of items using dual criteria of CVR and CVI, ensuring that the retained items were both relevant and essential. When the same 34 items were administered to students, item analysis independently highlighted the same nine underperforming items — those with low item–total correlations, weak discrimination, or ceiling effects. The fact that both expert judgment and empirical data led to the same 25-item solution strengthens the validity argument. It demonstrates that the scale reflects both theoretical coverage of the critical thinking construct and empirical coherence when administered to the target population. This dual confirmation addresses a common limitation in CT measurement, where expert consensus and field testing often diverge (Polit & Beck, 2006; Romero Jeldres, 2023).

At the same time, the study revealed some of the enduring challenges in measuring critical thinking. Several items that experts deemed important did not perform well statistically in the pilot, highlighting the persistent gap between theoretical relevance and empirical functioning of CT measures. This difficulty stems from the inherently multidimensional nature of critical thinking, which includes both cognitive operations (interpretation, inference, analysis) and dispositional aspects (open-mindedness, truth-seeking). Items that capture dispositional elements, in particular, may be endorsed positively by most students, leading to ceiling effects and weak discrimination. This finding echoes international research where CT is acknowledged as a construct that resists reduction into single-dimensional survey measures (Facione, 1990; Dwyer et al., 2014).

Another methodological contribution of this study is the use of combined CVR and CVI criteria for content validation. Many validation studies rely solely on CVI, which measures item relevance, but this study added CVR to capture essentiality. This dual-filter mechanism provided a stricter standard for item retention and increased expert consensus (S-CVI/Ave improved to 0.874). By ensuring that retained items were not only relevant but also indispensable, the study enhanced the scale's content validity and offered a more rigorous foundation for subsequent psychometric testing.

The incorporation of Raven's Advanced Progressive Matrices (RAPM) further strengthens the methodological framework. The symmetric distribution of RAPM scores confirmed that the sample was cognitively balanced, supporting the feasibility of matched-pair sampling for future intervention studies. This attention to group equivalence is often overlooked in CT validation research but is crucial for establishing internal validity in experimental designs.

Taken together, these findings contribute to both theory and practice. For theory, they illustrate that a carefully designed Critical Thinking Scale for teacher education can be both conceptually rich and empirically robust. For practice, the validated item scale provides teacher education institutions with a reliable and contextually grounded tool to assess the critical thinking skills of pre-service teachers. This is particularly timely in the Indian context, where the National Education Policy (NEP) 2020 has emphasized the integration of higher-order skills into all levels of education, including teacher preparation. A validated scale offers not only a mechanism to measure these outcomes but also a means to monitor the effectiveness of curricular reforms.

Nonetheless, some limitations must be acknowledged. The pilot sample size of 89, while sufficient for item analysis and reliability checks, was inadequate for exploratory or confirmatory factor analysis. The study also relied on self-report measures, which may capture perceived dispositions rather than actual reasoning performance. These limitations suggest that future research should conduct large-scale testing with EFA and CFA, examine criterion validity by linking CT scores to observed teaching behaviours, and explore measurement invariance across gender and linguistic groups.

In summary, the study demonstrates that integrating expert content validation (CVR + CVI) with empirical pilot testing can produce a psychometrically sound Critical Thinking Scale. The reduction of items improved both expert consensus and reliability, confirming that refinement enhances rather than weakens instrument strength. The final item scale provides a strong platform for further large-scale validation and offers teacher education institutions a valuable tool for aligning pedagogy with policy imperatives in fostering critical thinking.

6. Conclusion

This study sought to develop and validate a CTS for B.Ed. students by combining rigorous expert-based content validation with empirical pilot testing. The research process began with 34 items and, through the application of CVR and CVI indices, was refined to 25 items endorsed as both relevant and essential. Pilot testing with 89 students confirmed this reduction, with item analysis removing poorly performing items and reliability improving from an already acceptable $\alpha = 0.78$ to $\alpha = 0.80$ in the final version. The convergence of expert judgment and empirical evidence provides robust support for the validity of the 25-item scale.

The methodological contribution of this study lies in the dual use of CVR and CVI for content validation, which ensured stricter standards for item retention than either index alone. The subsequent pilot testing demonstrated that careful item refinement not only streamlines the instrument but also enhances reliability. The inclusion of Raven's Advanced Progressive Matrices (RAPM) to establish baseline cognitive equivalence adds further rigor, providing a foundation for matched-pair experimental designs in future research.

Practically, the validated scale offers teacher education programs a contextually relevant tool to assess and monitor CTS among pre-service teachers. Such assessment is critical in light of India's National Education Policy (NEP 2020), which emphasizes higher-order thinking as a core graduate attribute. The instrument can support diagnostic assessments at entry, formative evaluation of interventions, and summative judgments about program effectiveness.

At the same time, limitations of the current study must be acknowledged. The pilot sample, while adequate for item analysis, was insufficient for factor analysis. The single-institution context and reliance on self-report also limit generalizability. Future research should therefore conduct large-scale validation with Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA), test criterion-related validity by linking CT scores with teaching performance, and examine measurement invariance across subgroups. Integrating performance-based tasks alongside self-report measures would further strengthen the scale's ability to capture the multifaceted nature of critical thinking.

An important contribution of this study lies in its comparability with internationally recognized critical thinking instruments. Established tools such as the California Critical Thinking Skills Test (CCTST) and the Critical Thinking Self-Assessment Scale (CTSA) have been widely used across higher education contexts, yet they are often criticized for being commercially restricted, culturally biased, and costly to administer (Yuan et al., 2023; Saavedra & Opfer, 2012). In contrast, the 25-item CT scale developed in this study is openly available, contextually grounded in teacher education, and validated using both content indices (CVR, CVI) and pilot psychometrics. While the CCTST emphasizes cognitive reasoning skills through performance tasks, and the CTSA focuses on dispositions through self-report, the present instrument combines elements of both by addressing cognitive and dispositional dimensions within a single scale. This integrative approach enhances its applicability to pre-service teacher education, where the ability to think critically must be demonstrated not only in problem-solving tasks but also in reflective dispositions that inform teaching practice. Thus, the scale developed here complements and extends the international CT measurement landscape by offering a reliable, low-cost, and policy-relevant alternative for use in non-Western educational settings.

The future research directions related to CTS are listed as follows:

- Validate the CTS using Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) with larger samples.
- Test criterion validity by linking CTS scores with observed teaching performance.
- Examine measurement invariance across gender, language, and institutional contexts.

- Adapt and test the CTS in other teacher education programs across South Asia.
- Incorporate performance-based tasks alongside self-report items to capture critical thinking more comprehensively.
- Explore longitudinal use of the CTS to track CT development across teacher training.

In conclusion, this study has delivered a 25-item, reliable, and expert-endorsed Critical Thinking Scale for B.Ed. students. By addressing long-standing methodological gaps in CT measurement and aligning with national policy imperatives, the scale represents a valuable resource for both researchers and practitioners. With further validation, it has the potential to become a benchmark tool for assessing critical thinking in teacher education across diverse contexts.

Declarations

Source of Funding

This study received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing Interests Statement

The authors declare that they have no competing interests related to this work.

Consent for publication

The authors declare that they consented to the publication of this study.

Authors' contributions

Both the authors took part in literature review, analysis, and manuscript writing equally.

Availability of data and materials

Supplementary information is available from the authors upon reasonable request.

Ethical Approval

Ethical approval was obtained from the Institutional Research and Ethics Committee of the affiliated teacher education institution.

Institutional Review Board Statement

Not applicable for this study.

Informed Consent

Informed consent was obtained from all participants, who were assured of confidentiality, anonymity, and the voluntary nature of their involvement.

References

Ayre, C., & Scally, A.J. (2014). Critical values for Lawshe's content validity ratio: Revisiting the original methods of calculation. Measurement and Evaluation in Counseling and Development, 47(1): 79–86. https://doi.org/10.1177/0748175613513808.

Boateng, G.O., Neilands, T.B., Frongillo, E.A., Melgar-Quiñonez, H.R., & Young, S.L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. Frontiers in Public Health, 6: 149. https://doi.org/10.3389/fpubh.2018.00149.

Brown, T.A. (2015). Confirmatory factor analysis for applied research (2nd Edition). Guilford Press.

Dwyer, C.P., Hogan, M.J., & Stewart, I. (2014). An integrated critical thinking framework for the 21st century. Thinking Skills and Creativity, 12: 43–52. https://doi.org/10.1016/j.tsc.2013.12.004.

Ennis, R.H. (2011). The nature of critical thinking: An outline of critical thinking dispositions and abilities. University of Illinois. Retrieved from https://education.illinois.edu/docs/default-source/faculty-documents/robert-ennis/thenatureofcriticalthinking.

Facione, P.A. (1990). Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction (The Delphi Report). American Philosophical Association. Retrieved from https://eric.ed.gov/?id=ed315423.

Fabio, R.A., Romano, R., & Rizzo, S. (2025). Psychometric properties and validation of the Critical Reasoning Assessment (CRA). Personality and Individual Differences, 220: 113342. https://doi.org/10.1016/j.paid.2025.113344.

Galindo-Domínguez, H., Bezanilla, M.J., Poblete, M., & Fernández-Nogueira, D. (2023). A teachers' based approach to assessing the perception of critical thinking in university students. Frontiers in Education, 8: 1127705. https://doi.org/10.3389/feduc.2023.1127705.

Government of India (2020). National Education Policy 2020. Ministry of Education. Retrieved from https://www.education.gov.in/sites/upload_files/mhrd/files/nep_final_english.pdf.

Hultquist, T.B., Milner, R., & Taylor, J.M. (2023). Refinement and evaluation of the Critical Thinking Self-Assessment Scale (CTSAS). Journal of Nursing Measurement, 31(2): 393–405. https://doi.org/10.1891/jnm-2024-0061.

Lawshe, C.H. (1975). A quantitative approach to content validity. Personnel Psychology, 28(4): 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x.

Lynn, M.R. (1986). Determination and quantification of content validity. Nursing Research, 35(6): 382–385. https://doi.org/10.1097/00006199-198611000-00017.

Messick, S. (1994). The validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50(9): 741–749. https://doi.org/10.1037/0003-066x.50.9.741.

Nunnally, J.C., & Bernstein, I.H. (1994). Psychometric theory (3rd Edition). McGraw-Hill.

Önal, İ. (2025). Building critical thinking in pre-service teachers: Development and validation of a scale. Education and Training, 67(1): 45–62. https://doi.org/10.1371/journal.pone.0330536.

Payan-Carreira, R., Gomes, A., & Manuel, S. (2022). Development and validation of a short form of the Critical Thinking Self-Assessment Scale. Education Sciences, 12(10): 676. https://doi.org/10.3390/educsci12100676.

Polit, D.F., & Beck, C.T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. Research in Nursing & Health, 29(5): 489–497. https://doi.org/10.1002/nur.20147.

Rodríguez-Rojas, M.P. (2024). Critical Thinking Evaluation Scale (CTES): Design and validation in Colombia. SAGE Open, 14(2): 21582440241297418. https://doi.org/10.1177/21582440241297418.

Romero Jeldres, M. (2023). A review of Lawshe's method for calculating content validity in the social sciences. Frontiers in Education, 8: 1136552. https://doi.org/10.3389/feduc.2023.1136552.

Saavedra, A.R., & Opfer, V.D. (2012). Learning 21st-century skills requires 21st-century teaching. Phi Delta Kappan, 94(2): 8–13. https://doi.org/10.1177/003172171209400203.

Tiruneh, D.T., Verburgh, A., & Elen, J. (2017). Effectiveness of critical thinking instruction in higher education: A systematic review of intervention studies. Higher Education Studies, 7(1): 1–17. https://doi.org/10.5539/hes.v7n1p1.

Worthington, R.L., & Whittaker, T.A. (2006). Scale development research: A content analysis and recommendations for best practices. The Counseling Psychologist, 34(6): 806–838. https://doi.org/10.1177/0011000006288127.

Yuan, R., Zheng, X., & Zhang, L. (2023). Re-examining critical thinking in teacher education: A review of the literature. Teaching and Teacher Education, 125: 104002. https://doi.org/10.1016/j.tate.2023.104002.