International Journal of Learning, Teaching and Educational Research
Vol. 25, No. 8, pp. 968-985, August 2026
https://doi.org/10.26803/ijlter.25.8.38
Received Jun 30, 2026; Revised Aug 4, 2026; Accepted Aug 12, 2026

TRACE for Evidence-to-Claim Coherence in Undergraduate Thesis Writing: Development and Descriptive Evaluation

Luijim Jose*
Nueva Ecija University of Science and Technology
Cabanatuan City, Nueva Ecija, Philippines

Jacqueline G. Gumallaoi, Jonalyn Sad-ayan-Lacambra, Jessica E. Ayawan, Imelda N. Binay-an, Cynthia M. Carino and Olga S. Tresmanio
Ilocos Sur Polytechnic State College
Ilocos Sur, Philippines

Abstract. Evidence-to-claim coherence requires thesis conclusions and recommendations to remain traceable to research questions, analytical results, evidence, warrants, and methodological boundaries. Grounded in scholarship on argumentation, validity, feedback literacy, and rubric assessment, this study developed the Thesis Reasoning and Argument Coherence Evaluation Framework (TRACE) for undergraduate thesis writing. TRACE integrates an Evidence–Claim Traceability Matrix, a six-dimensional Claim Calibration and Coherence Rubric, a Taxonomy of Research Inference Errors, and a secondary Argument Coherence Index (ACI). Framework development combined conceptual synthesis, exploratory examination of 42 anonymized thesis chapter sets, expert-guided refinement, and rater calibration. A subsequent descriptive evaluation involved 120 third-year undergraduates in four intact thesis-writing sections at a public teacher-education institution in Central Luzon, Philippines. Two sections received TRACE-supported instruction; two received equal-duration conventional thesis-writing instruction without TRACE tools. Ten doctoral-level experts in research methodology, academic writing, educational assessment, instrument development, curriculum, measurement, or thesis supervision appraised high-level framework statements (S-CVI/Ave = .92). Three trained raters scored baseline drafts and final outputs; average-measure absolute-agreement coefficients were ICC (2,3) = .854 and .883, respectively. All sections had higher final means, but one comparison section showed a trajectory comparable to that of the TRACE sections. Very low internal consistency among final dimension scores (α = .09) argued against a homogeneous-scale interpretation of the ACI. TRACE is recommended

Citation:
Jose, L., Gumallaoi, J, G., Sad-ayan-Lacambra, J., Ayawan, J, E., Binay-an, I, N., Carino, C, M., & Tresmanio, O, S. (2026). TRACE for Evidence-to-Claim Coherence in Undergraduate Thesis Writing: Development and Descriptive Evaluation. *International Journal of Learning, Teaching and Educational Research*, 25(8), 968–985. https://doi.org/10.26803/ijlter.25.8.38

*Corresponding author: Luijim Jose; *luijimjose@ineust.ph.education*

for cautious formative and diagnostic use, with the dimension profile primary and the ACI provisional.

Keywords: undergraduate research; evidence-to-claim coherence; claim calibration; inference errors; Philippines

1. Introduction

Research credibility depends on whether claims, conclusions, and recommendations are proportionate to the evidence. An analytical result does not automatically justify every interpretation. Researchers must show how the result answers the research question, why it supports a claim, how strongly the claim may be stated, and what limits follow from the design, sample, measurement, analysis, and context. Toulmin's (2003) model locates the warrant between data and claim and uses qualifiers and rebuttals to delimit defensible assertions. Contemporary argumentation scholarship likewise emphasizes the epistemic standards and reliable processes for evaluating evidence (Duncan & Chinn, 2025).

Evidence-to-claim coherence is especially consequential in undergraduate theses. Students may infer causality from a nonexperimental design, equate statistical significance with practical importance, treat nonsignificance as proof of no relationship, extend a contextual qualitative interpretation beyond its evidentiary reach, introduce conclusions absent from the findings, or recommend actions unsupported by the study. Quantitative interpretation must consider design, uncertainty, effect magnitude, measurement, and alternative explanations (Greenland et al., 2016; McShane et al., 2024; Rutkowski et al., 2024). Qualitative interpretation requires congruence among the question, methodological orientation, empirical materials, analysis, reflexivity, and context (Braun & Clarke, 2025; Levitt et al., 2018). In this study, methodological boundaries are the design-, evidence-, analysis-, and context-based conditions that limit certainty, causal force, generalizability, and practical implications (American Educational Research Association et al., 2014; Kane, 2013).

Thesis writing is therefore a disciplinary and methodological practice, not merely a language exercise. Research on thesis writing emphasizes genre specificity, supervision, textual practice, and iterative revision (Flowerdew & Petrić, 2024; Grohnert et al., 2024; Jusslin & Hilli, 2024; Lymer et al., 2024). Feedback becomes consequential when writers understand criteria and revise the reasoning underlying their claims, rather than only surface features (Curtis et al., 2025; Quinlan & Pitt, 2025; Teng & Ma, 2024). Analytic rubrics can support that work when constructs are explicit, raters are prepared, and scores remain within the available validation evidence (Arias-Hermoso et al., 2025; Harsch et al., 2024; Jönsson & Svingby, 2007).

The Research Coherence Alignment Framework (RCAF) addresses prospective alignment among questions, constructs, evidence sources, instruments, and analyses during proposal design (Jose, 2026). TRACE addresses the later interpretive stage: the progression from research question to analytical result, evidence, warrant, methodological boundary, calibrated claim, conclusion, and

recommendation. Its four components are separate pathway construction, multidimensional evaluation, diagnosis, and descriptive summarization. Because the present classroom phase involved only four intact sections and non-equivalent baseline and final artifacts, its results are section-specific descriptive trajectories rather than estimates of instructional effectiveness.

1.1 Statement of the Research Problem and Research Gaps

Undergraduate writers may complete appropriate data collection and analysis, yet produce interpretations that exceed or detach from the evidence. Common feedback such as "support this claim," "do not overgeneralize," or "align the recommendation" identifies a symptom but not whether the breakdown concerns evidence, warrant, causal language, certainty, scope, or recommendation traceability. Uncorrected errors weaken the validity of individual theses and, when repeated across programs, can undermine institutional research credibility, research integrity, and evidence-informed decision-making.

Five gaps motivated TRACE. Conceptually, the complete pathway from question to recommendation is rarely represented as one system. Methodologically, a cross-method framework must preserve the distinct evidentiary warrants of quantitative, qualitative, mixed-methods, action research, and textual inquiry. Diagnostically, broad labels such as "weak argument" do not distinguish among causal overreach, significance inflation, literature substitution, or recommendation detachment. Pedagogically, writers need formative construction and revision procedures, not only summative judgments. Finally, validation and implementation evidence are required before a new framework is used broadly or for high-stakes decisions.

The exploratory review of 42 chapter sets informed construct development but did not estimate the prevalence of errors. The corpus was neither representative nor coded using a prospectively validated protocol by independent prevalence coders. Reporting a percentage would therefore create unwarranted precision; prevalence should be established in a separate, adequately sampled study.

1.2 Research Objectives

This study aimed to develop and descriptively evaluate TRACE as an instructional and diagnostic system for undergraduate thesis writing. Specifically, it sought to:

1. conceptualize and operationalize the TRACE pathway, four operational components, and six rubric dimensions;
2. obtain preliminary expert content-relevance evidence for the high-level TRACE architecture and for the dedicated high-level TRIE taxonomy statement, and document expert-guided refinements;
3. estimate the dependability of three-rater baseline and final-output scores and examine implications of the interdimension pattern for the secondary ACI; and
4. describe section-specific baseline-to-final differences across the six dimensions and secondary ACI summaries.

The longer-term instructional goal is to help writers produce claims, conclusions, and recommendations proportionate to the evidence. The diagnostic goal is to help writers, supervisors, and raters identify breakdowns within the evidence-to-claim pathway.

1.3 Research Questions

1. How was TRACE conceptualized and operationalized as an instructional and diagnostic system for undergraduate thesis writing?
2. What preliminary expert content-relevance evidence was obtained for the high-level TRACE architecture, including TRIE, and how did expert feedback refine the framework?
3. How dependable were the three-rater scores, and what did the interdimension pattern indicate about the secondary ACI?
4. What section-specific descriptive differences occurred from baseline to final output across the six rubric dimensions and ACI?

1.4 Significance and Contribution of the Study

TRACE contributes conceptually by defining evidence-to-claim coherence as a complete, traceable pathway rather than a general quality of academic writing. It contributes methodologically by retaining common reasoning questions while allowing evidence and warrants to vary across research traditions. Its diagnostic contribution is TRIE's location of consequential errors within a manuscript pathway, rather than within a generic list of formal or informal fallacies. Practically, the ECTM supports construction, the CCCR supports multidimensional formative evaluation, TRIE supports targeted diagnosis, and the ACI provides only a secondary summary.

These functions are relevant beyond one institution because researchers in different settings must connect evidence, warrants, claims, conclusions, and recommendations. The present findings do not establish universal applicability; they provide an architecture and preliminary evidence that other disciplines, institutions, and countries can adapt and independently validate.

2. Literature Review

2.1 Evidence, Warrants, and Claim Calibration

Argumentation requires more than placing evidence beside a conclusion. Toulmin (2003) distinguishes claims, data, warrants, backing, qualifiers, and rebuttals, while current work examines the epistemic quality of the reasoning processes that connect them (Duncan & Chinn, 2025). Explicit evidence-reasoning instruction can help learners distinguish observations from inferences and justify claims (Du & List, 2024). The Data Description, Claim, Evidence, and Reasoning framework applies a related sequence in undergraduate science instruction (Browne et al., 2026).

Within TRACE, claim calibration means matching certainty, causal language, scope, and practical implications to the evidence. In quantitative inquiry, calibration depends on design adequacy, uncertainty, effect magnitude, measurement, assumptions, and alternative explanations; statistical adjustment

cannot create causal identification that the design lacks (Greenland et al., 2016; McShane et al., 2024; Rutkowski et al., 2024). In qualitative inquiry, it depends on methodological congruence, empirical grounding, reflexivity, and contextual reach (Braun & Clarke, 2025; Levitt et al., 2018). Mixed-methods claims must reflect the quality and purpose of integration; action-research claims should distinguish local improvement from general effects; textual claims must remain tied to the corpus, lens, and context.

2.2 Thesis Writing, Disciplinary Variation, and Supervision

Writers must select relevant findings, interpret their significance, relate them to scholarship, state limitations, and formulate traceable conclusions and recommendations. Thesis-writing studies emphasize genre, supervision, and iterative textual practice (Flowerdew & Petrić, 2024; Jusslin & Hilli, 2024; Lymer et al., 2024). Difficulties include selecting key findings, organizing results, and understanding the purposes of sections (Chien & Li, 2024). Effective supervision depends on the interaction among student characteristics, supervisory action, relationships, and expected outcomes (Grohnert et al., 2024), while structured process models can make thesis development more visible (Praestegaard Larsen, 2025).

Calibration also varies by discipline. Writers construct value arguments, cohesion, and stance differently across fields (Golparvar et al., 2025; Hu & Bonsu, 2025; Kurt & Kafes, 2025). An experimental warrant may require causal identification and control of alternatives; an interpretive warrant may require reflexivity, depth, contextual grounding, and congruence. TRACE, therefore, asks common questions but evaluates answers against the relevant methodological tradition.

2.3 Feedback Literacy and Formative Rubric Use

Feedback contributes to learning when students can interpret standards and act on evaluative information (Quinlan & Pitt, 2025). Exemplars, comparison, self-assessment, and metacognitive regulation can develop student feedback literacy (Curtis et al., 2025; Teng & Ma, 2024). Feedback literacy also depends on teachers and institutions, whose practices are shaped by workload, relationships, and organizational constraints (Musaeus et al., 2026).

Summative rubrics grade or classify performance; formative rubrics clarify expectations, structure dialogue, and guide revision. Analytic rubrics can improve transparency and consistency, but reliability alone does not establish valid interpretation or use (Harsch et al., 2024; Jönsson & Svingby, 2007; Kane, 2013). Recent work supports multidimensional rubrics, criteria-referenced tools, rubric co-creation, and attention to how feedback design directs learners' focus (Arias-Hermoso et al., 2025; Grainger et al., 2025; Panadero et al., 2025; Yan, 2024). TRACE is therefore formative and diagnostic; current evidence does not support its use for thesis approval, ranking, or institutional proficiency classification.

2.4 Research Inference Errors and Rationale for TRACE

Research inference errors occur when a claim violates the logical or methodological conditions required by the evidence. Literature substitution, for example, occurs when an external source is used as a replacement for evidence

that the present study should have generated. Prior literature may compare or explain findings, but it cannot establish an unobserved condition among the present participants. Other TRIE categories include question-to-finding displacement, evidence-to-claim overreach, correlation-to-causation conversion, significance inflation, nonsignificance absolutism, theme–evidence disconnection, qualitative generalization error, conclusion introduction, and recommendation detachment.

TRIE differs from a generic fallacy list because its categories are located within a research-question-to-recommendation pathway and are judged against the design, evidence, analysis, and methodological tradition. Existing scholarship supports individual aspects of TRACE, but the reviewed literature did not identify a thesis-oriented, cross-method system that combines traceability, method-sensitive warrants, explicit boundaries, claim calibration, multidimensional evaluation, and inference-error diagnosis. TRACE's originality lies in this coordinated integration, not in claiming that its individual concepts are new.

3. Conceptual Framework

TRACE defines evidence-to-claim coherence as the degree to which a claim answers a research question, follows from an analytical result, is supported by relevant evidence, is connected through a defensible warrant, respects methodological boundaries, is calibrated to the evidence, and leads to a congruent conclusion or recommendation. The pathway is recursive rather than mechanically linear: writers may revisit analyses, narrow claims, or revise recommendations.

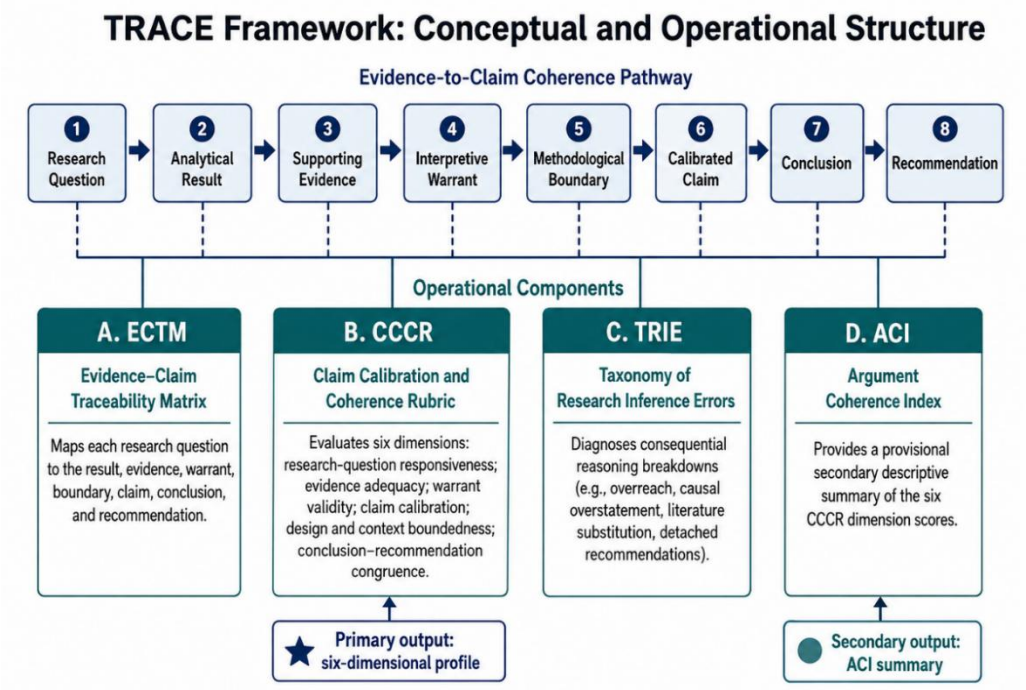

Figure 1: Conceptual and operational structure of TRACE

Note. ECTM = Evidence–Claim Traceability Matrix; CCCR = Claim Calibration and Coherence Rubric; TRIE = Taxonomy of Research Inference Errors; ACI = Argument Coherence Index.

3.1 Operational Components

Table 1: Operational components of TRACE

Component	Primary function	Intended formative use
ECTM	Maps each question to result, evidence, warrant, boundary, claim, conclusion, and recommendation.	Construction, self-examination, and revision.
CCCR	Evaluates six dimensions using four substantive performance anchors.	Profile of comparative strengths and weaknesses.
TRIE	Diagnoses consequential reasoning breakdowns within the pathway.	Specific language for targeted feedback and revision.
ACI	Sums the six dimension means using provisional equal weighting.	Secondary summary interpreted with the profile and diagnoses.

The CCCR dimensions are research-question responsiveness, evidence adequacy, warrant validity, claim calibration, design and context boundedness, and conclusion–recommendation congruence. Each is scored from 1 (weak) to 4 (strong), with prospectively authorized quarter-point interpolation and written justification. The six-dimension profile is primary. Their sum yields an ACI from 6 to 24, but no cut scores, proficiency bands, minimally important differences, or high-stakes uses are proposed. Because the components are distinct rather than interchangeable indicators of a single trait, low internal consistency does not invalidate the profile; however, it does restrict the interpretation of the total (Bollen & Lennox, 1991; Diamantopoulos & Winklhofer, 2001).

4. Methodology

This study used a methodological framework-development design followed by a descriptive, nonrandomized evaluation across four intact undergraduate thesis-writing sections.

4.1 Framework Development

Development combined conceptual synthesis and an exploratory review of 42 anonymized chapter sets from five undergraduate teacher-education programs produced from Academic Years 2022–2023 to 2024–2025. The corpus included 24 quantitative, 10 qualitative, six mixed-methods, and two action-research sets and was separate from the classroom sample. A structured guide examined each question, result, evidence, warrant, boundary, claim, conclusion, and recommendation. Two team members first examined 12 sets to refine the guide; the remaining 30 were reviewed with the revision, and six were independently re-examined to clarify provisional categories. Disagreements informed conceptual refinement, not prevalence estimates or formal coder reliability.

4.2 Context and Participants

The classroom evaluation occurred during the second semester of Academic Year 2025–2026 in four regularly formed third-year thesis-writing sections at a public teacher-education institution in Central Luzon, Philippines. Condition allocation was based on timetable feasibility before baseline scores were examined. Of 124

enrolled students, 122 consented; one lacked an eligible baseline draft, and one withdrew because of prolonged absence. The complete-case sample comprised 120 students, with 30 per section and no imputation. Participants had a mean age of 21.6 years (SD = 1.0, range = 20–25); 84 were women and 36 men. Manuscripts comprised 68 quantitative, 30 qualitative, 16 mixed-methods, four action-research, and two textual or documentary studies.

Table 2: **Section-level participant and implementation characteristics**

Characteristic	C1 TRACE	C2 TRACE	C3 Comparison	C4 Comparison
Analyzed n	30	30	30	30
Age, M (SD)	21.7 (1.0)	21.5 (0.9)	21.6 (1.1)	21.6 (1.0)
Women/men	22/8	21/9	20/10	21/9
Baseline ACI, M	13.37	13.18	13.58	12.02
Attendance	94.3%	93.7%	93.9%	93.3%
Adviser consultations, M	3.8	3.7	3.9	3.8

Note. ACI = Argument Coherence Index. Consultation counts did not capture duration, content, quality, or influence. C4's lower baseline mean is considered in interpreting its trajectory.

4.3 Instructional Conditions and Fidelity

The same instructor taught all sections for equal scheduled contact time. TRACE instruction comprised five 90-minute group sessions and one 30-minute individual consultation (480 minutes) covering evidence, warrants, ECTM mapping, TRIE diagnosis, CCCR-guided evaluation, and integrated revision. The active comparison condition covered the presentation of findings, literature integration, cohesion, claim clarity, conclusions, recommendations, and general manuscript feedback, but excluded TRACE terminology and tools.

The implementation checklist contained 15 planned activities per TRACE section. C1 completed all 15; C2 completed 14 because a second peer-review exercise was omitted after a suspended meeting. Core activities were completed in both sections. Fidelity evidence came from attendance, class records, and an instructor-completed checklist, not independent observation. Following implementation reporting principles (Baelen et al., 2023), the reported 96.7% therefore represents documented activity completion, not audited delivery quality or participant uptake. No formal contamination survey was administered.

4.4 Measures, Expert Appraisal, and Rater Preparation

Raters examined every question-to-recommendation pathway and assigned one holistic score to each CCCR dimension. Missing required elements affected the relevant score. Dimension means were averaged across three raters, and the ACI summed the six unrounded means. The baseline artifact was a pre-instruction thesis draft; the final artifact was the completed output. Comparing these artifacts was justified as an examination of authentic thesis-development trajectories rather than change on parallel assessment tasks. Final outputs had received additional instruction, feedback, consultation, reading, and revision; therefore, differences cannot be attributed exclusively to TRACE. Files were de-identified and randomized separately for each rater; at least 52 days separated baseline and

final-rating periods. Complete masking of the occasion was impossible because final manuscripts were generally more complete.

Ten independent doctoral-level experts with 8–26 years of relevant experience appraised 10 high-level statements on a four-point relevance scale. The panel represented research methodology (3), applied linguistics and academic writing (2), educational assessment and instrument development (2), curriculum and instruction (1), educational measurement and statistics (1), and graduate research supervision (1). I-CVI, S-CVI/Ave, S-CVI/UA, and modified kappa were descriptive indicators (Polit & Beck, 2006; Polit et al., 2007). The appraisal addressed high-level relevance rather than comprehensive validation of every field, descriptor, category, scoring rule, or use.

Three independent raters held doctorates in educational management, applied linguistics, or curriculum and instruction, and had 9–15 years of relevant experience. An eight-hour program across four sessions covered construct orientation, guided and independent scoring, method-sensitive examples, interpolation, written justifications, and TRIE application. Formal scoring began after 88.9% exact-or-adjacent agreement in practice ratings. Inter-rater dependability used two-way random-effects, absolute-agreement ICCs; ICC(2,3) represented the three-rater mean and ICC (2,1) a single rater (Koo & Li, 2016; McGraw & Wong, 1996). Ten-thousand-replicate participant-resampling bootstrap intervals preserved each manuscript's three ratings.

4.5 Data Analysis, Bias, and Ethics

Expert results were summarized using content validity indices and modified kappa. Dependability analyses included single- and average-measure ICCs, dimension-level coefficients, and rater-specific difference-score ICCs. Coefficient alpha and interdimension correlations were examined only as score-structure diagnostics. Baseline means, final means, and differences were calculated separately for each section and dimension. No pooled condition contrast, student-level significance test, or interval was interpreted as an instructional effect because four assignment units cannot support stable cluster-adjusted inference (McNeish & Stapleton, 2016; Qiu et al., 2024). Analyses used R 4.4.2.

Threats included nonrandom allocation, four sections, section composition, schedule, peer interaction, adviser support, topic complexity, ordinary thesis maturation, non-equivalent artifacts, regression toward the mean, attrition, contamination, researcher allegiance, instructor-reported fidelity, and alignment between instruction and the CCCR. Independent raters and the separation of research ratings from grades reduced, but did not eliminate, allegiance concerns. Consent was obtained by a research assistant with no teaching, advising, grading, or scoring role; participation did not affect grades or support, and research ratings were withheld from instructors and advisers until final grades were submitted.

5. Results and Findings

The results are organized according to the four research questions.

5.1 Conceptualization and Operationalization of TRACE

Conceptual synthesis established the question-to-recommendation pathway; the 42-set review identified recurrent breakdowns and informed preliminary ECTM fields, CCCR dimensions, and TRIE categories. Expert feedback and rater preparation refined terminology, method-sensitive examples, descriptor boundaries, interpolation guidance, and restrictions on ACI interpretation.

Table 3: Development issues and corresponding revisions

Issue identified	Revision incorporated
Evidence adequacy overlapped with warrant validity.	Evidence was limited to relevance and sufficiency; warrant validity addressed the defensibility of the reasoning link.
Methodological limits were insufficiently visible.	A boundary field was strengthened in the ECTM and represented in the CCCR.
Descriptors and examples were too general or quantitatively weighted.	Observable anchors and qualitative, mixed-methods, action-research, and textual examples were added.
The ACI could conceal serious weaknesses.	The complete profile and consequential TRIE diagnoses became mandatory companions to the ACI.
Classification bands resembled validated cut scores.	The bands were removed; no proficiency categories were proposed.

These results answer Research Question 1: TRACE was operationalized as an integrated system for pathway construction, multidimensional evaluation, inference-error diagnosis, and formative summarization, rather than as a general writing checklist or single-score instrument.

5.2 Preliminary Expert Appraisal

All 10 experts endorsed construct clarity and pathway coherence (I-CVI = 1.00); nine endorsed each of the remaining eight statements (I-CVI = .90). The S-CVI/Ave was .92, the S-CVI/UA was .20, and modified kappas ranged from .90 to 1.00.

Table 4: Item-level expert appraisal of TRACE

Appraisal statement	Endorsing	I-CVI	Kappa
Construct clarity	10/10	1.00	1.00
Pathway coherence	10/10	1.00	1.00
ECTM relevance	9/10	.90	.90
CCCR relevance	9/10	.90	.90
TRIE taxonomy adequacy	9/10	.90	.90
ACI as a formative secondary summary	9/10	.90	.90
Instructional usefulness	9/10	.90	.90
Diagnostic usefulness	9/10	.90	.90
Cross-method applicability	9/10	.90	.90
Overall methodological and diagnostic relevance	9/10	.90	.90

Note. Ratings of 3 or 4 were classified as endorsements. The appraisal supports the high-level architecture but not the comprehensive validation of every TRACE element.

These results answer Research Question 2 by showing favorable preliminary content relevance, including endorsement of the high-level TRIE statement, while preserving appropriate limits on the evidence.

5.3 Rater Dependability and Score Structure

Baseline scores yielded ICC (2,1) = .661 and ICC(2,3) = .854, 95% bootstrap CI [.801, .891]. Final scores yielded ICC (2,1) = .715 and ICC (2,3) = .883, 95% CI [.830, .916]. Dimension-level ICC (2,3) s ranged from .811 to .868 at baseline and .850 to .907 at final output; difference-score ICC (2,3) was .789, 95% CI [.711, .847].

Table 5: Inter-rater dependability of CCCR scores

Score	Baseline ICC (2,3)	Final ICC (2,3)
Research-question responsiveness	.842	.886
Evidence adequacy	.868	.907
Warrant validity	.824	.869
Claim calibration	.811	.850
Design/context boundedness	.856	.903
Conclusion–recommendation congruence	.838	.867
ACI (three-rater mean)	.854 [.801, .891]	.883 [.830, .916]

The final dimension scores had α = .09 and an average interdimension correlation of .016 (range = −.18 to .21), arguing against a homogeneous reflective-scale interpretation and leaving equal weighting unvalidated. Three anonymized excerpts from the raters' written justifications illustrate the diagnostic reasoning:

> *"The reported association answers the question, but the manuscript states that the program caused the outcome. Because the design is correlational, the claim should indicate an association rather than an effect."*

> *"Participant accounts support the theme, but the claim is extended to all preservice teachers. The interpretation should be restricted to the participants and context examined."*

> *"The recommendation for institution-wide implementation is not supported by evidence of effectiveness, feasibility, or acceptability. It should be reframed as a recommendation for pilot testing."*

These results answer Research Question 3: three-rater means were more dependable than single-rater scores, but the profile—not a unitary ACI—should guide interpretation.

5.4 Section-Specific Descriptive Trajectories

Every section had higher final means across all six dimensions, but the magnitude of these differences varied within and between conditions.

Table 6: Section-level baseline-to-final differences

Section	Condition	RQR	EA	WV	CC	DCB	CRC	ΔACI
C1	TRACE	0.55	0.65	0.59	0.62	0.61	0.54	3.56
C2	TRACE	0.53	0.60	0.56	0.61	0.56	0.47	3.33
C3	Comp.	0.43	0.43	0.41	0.45	0.42	0.45	2.59
C4	Comp.	0.56	0.59	0.56	0.61	0.57	0.49	3.38

Note. Comp. = active comparison; RQR = research-question responsiveness; EA = evidence adequacy; WV = warrant validity; CC = claim calibration; DCB = design/context boundedness; CRC = conclusion–recommendation congruence; ΔACI = ACI difference.

Table 7: Section-level ACI summaries

Section	Condition	Baseline M (SD)	Final M (SD)	Difference M (SD)	95% CI
C1	TRACE	13.37 (2.55)	16.93 (2.39)	3.56 (2.58)	[2.60, 4.53]
C2	TRACE	13.18 (2.09)	16.51 (1.53)	3.33 (2.15)	[2.52, 4.13]
C3	Comparison	13.58 (2.77)	16.17 (2.35)	2.59 (2.98)	[1.48, 3.70]
C4	Comparison	12.02 (2.50)	15.40 (2.39)	3.38 (2.48)	[2.45, 4.31]

Note. Confidence intervals describe student-level differences within sections and are not cluster-adjusted instructional-effect intervals.

Dimension differences ranged from 0.41 to 0.65 on the four-point scale. Substantively, these movements indicate more explicit responsiveness to questions, stronger evidence and warrants, better-calibrated claims, clearer acknowledgment of methodological boundaries, and improved traceability of conclusions and recommendations. They should not, however, be labeled small, moderate, or large because TRACE has no validated minimally important difference or responsiveness threshold. C4 began at 12.02, 1.56 points below C3, and therefore had more numerical room for upward movement; despite a larger difference, its final mean remained 0.77 points below C3.

Plausible, nonexclusive explanations include baseline position, regression toward the mean (Barnett et al., 2005), active comparison instruction, adviser support, feedback, revision time, peer interaction, engagement, topic complexity, and possible circulation of strategies. No dimension displayed a pattern unique to TRACE. These results answer Research Question 4 without establishing causation.

6. Discussion

6.1 Contributions and Measurement Interpretation

TRACE's novelty lies in integrating six non-substitutable requirements into a single thesis-oriented pathway and four coordinated tools. General writing rubrics may evaluate organization, language, source use, or argument quality; TRACE asks whether a research claim can be traced backward to its question, result, evidence, warrant, and boundary and forward to a conclusion and recommendation. Claim calibration is distinctive because a plausible claim can still be expressed with greater certainty, causal force, breadth, or practical reach than the design permits.

The expert results support continued development rather than comprehensive validation. Likewise, dependable scoring is not equivalent to construct validity or generalizability. The low alpha and near-zero average interdimension correlation reinforce the a priori view that the profile is primary. Equal totals may represent different weaknesses, and a seemingly acceptable ACI can conceal serious causal overreach or a detached recommendation. The ACI should not determine grades, thesis approval, graduation, publication readiness, or institutional rankings.

6.2 Classroom Patterns and Comparison Section C4

The positive trajectories are directionally consistent with interventions that make evidence-based reasoning explicit. DCER separates data description, claims, evidence, and reasoning in undergraduate science instruction (Browne et al., 2026), while Du and List (2024) directly teach learners to evaluate evidence and justify conclusions. TRACE extends this instructional logic from individual explanations to the complete thesis pathway, adding method-sensitive warrants, methodological boundaries, conclusion–recommendation congruence, and parallel error diagnosis. Consequential-feedback research likewise supports criteria that students can use for substantive revision rather than surface correction (Quinlan & Pitt, 2025).

Unlike stronger intervention studies, however, the present design was not randomized and contained only four assignment units. The active comparison sections also received instruction, consultation, feedback, and revision time. C4's low baseline and possible regression toward the mean contributed to numerical opportunity, while unmeasured differences in adviser support, peer processes, manuscript completeness, motivation, topic complexity, and strategy sharing remain plausible. C3 and C4 differed despite the same nominal condition, demonstrating why two sections should not be averaged and interpreted as a stable condition effect.

6.3 Diagnostic, Institutional, and Curricular Implications

TRIE can be applied in four steps: identify the claim or recommendation; trace it backward through the question, result, evidence, warrant, and boundary; determine whether a specific breakdown matches a TRIE category; and provide a revision direction that restores traceability and calibration. Recommendation detachment is especially consequential because a local descriptive finding may otherwise be converted into institution-wide policy without evidence of effectiveness, feasibility, cost, or acceptability.

Curricular integration should be staged. Introduce evidence, warrants, claims, and boundaries in research methods; use the ECTM during question and analysis planning; apply TRIE in method-sensitive interpretation workshops; and use the CCCR for self-assessment, peer review, and supervisor conferences. The ACI should be introduced only after the profile and its limits are understood. Full use may increase workload because instructors must review matrices, six dimensions, diagnoses, and revisions. Phased use at thesis milestones, shared calibration, exemplar libraries, and digital records can reduce the burden.

Institutional standardization should create a shared vocabulary and transparent documentation, not impose identical warrants across disciplines. Explicit criteria may help students question and understand judgments, but punitive use could reinforce supervisory control; TRACE should remain dialogic. Action research, textual, arts-based, Indigenous, participatory, and design-oriented inquiry requires adapted descriptors that respect situated, interpretive, relational, or creative warrants.

6.4 Limitations and Future Research

The principal limitations are four nonrandomized sections, one institution and instructor, non-equivalent artifacts, incomplete measurement of adviser support, instructor-reported fidelity, no formal contamination check, possible researcher allegiance, limited masking of occasion, high-level rather than component-level expert appraisal, and sparse representation of action-research and textual studies. Student and instructor perceptions were not measured and should be examined through dedicated instruments and research questions rather than inferred from scores.

Future effectiveness studies should use larger multisite samples, independently implemented active-control sections, random or defensibly matched cluster allocation, equivalent and delayed-transfer tasks, preregistered analyses, and external criteria. The present eight-hour rater-preparation program is a reasonable provisional minimum starting point, not a universal requirement; scoring should begin only after a preregistered agreement benchmark and should include scheduled recalibration. An independent fidelity protocol should use a session-specific checklist, external observation of randomly selected sessions in every section, attendance and dosage records, student-exposure measures, documented deviations, and adjudication of observer–instructor discrepancies. Contamination checks should ask students and instructors about shared files and terminology, record cross-section adviser assignments, and document exposure outside the assigned condition.

A digital TRACE platform could support pathway mapping, revision histories, scoring, and profile visualization. AI assistance may flag possible evidence–claim mismatches or generate questions for human review, but diagnoses must remain provisional and cannot independently determine thesis quality. Doctoral adaptation should require stronger evidence of originality, theoretical contribution, methodological sophistication, integration across studies, and disciplinary argumentation rather than merely relabeling undergraduate descriptors.

7. Conclusion

This study developed TRACE as an integrated system for constructing, evaluating, diagnosing, and revising evidence-to-claim reasoning in undergraduate theses. The four components operationalized a complete question-to-recommendation pathway; experts favorably endorsed the high-level architecture, including TRIE; and three-rater means showed stronger dependability than single-rater scores. All four sections had higher final means,

but one active comparison section had a trajectory comparable with the TRACE sections, and the dimension structure did not support a homogeneous-scale interpretation of the ACI.

Accordingly, TRACE's most defensible present use is cautious, formative, and diagnostic. The CCCR profile, written justifications, and consequential TRIE diagnoses should remain primary; the ACI should remain a provisional secondary summary without cut scores or high-stakes applications. Broader use requires independent multisite validation, stronger allocation, equivalent and transfer tasks, audited implementation, contamination checks, response-process evidence, and method-specific adaptation.

8. Conflict of Interest

The authors declare no conflicts of interest related to the research, authorship, or publication of this article.

9. Acknowledgments

The authors gratefully acknowledge the institutional support of Nueva Ecija University of Science and Technology and Ilocos Sur Polytechnic State College, as well as the experts, raters, research assistant, and participating undergraduate thesis writers.

The authors acknowledge the use of Grammarly and ChatGPT by OpenAI during manuscript revision. Grammarly supported grammar and sentence-level clarity. ChatGPT assisted with organizing reviewer comments, structural revision, selected language editing, and consistency checks. These tools did not generate or alter empirical data or statistical results. The authors verified the sources and technical information and retain responsibility for the manuscript's accuracy, originality, and integrity.

10. References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association. https://www.testingstandards.net/open-access-files.html

Arias-Hermoso, R., Garro Larrañaga, E., & Imaz Agirre, A. (2025). Assessing academic and disciplinary literacies: Rubric validation to measure argumentation, comparison and source-based writing skills. *Journal of Language and Education, 11*(3), 60–75. https://doi.org/10.17323/jle.2025.27560

Baelen, R. N., Gould, L. F., Felver, J. C., Schussler, D. L., & Greenberg, M. T. (2023). Implementation reporting recommendations for school-based mindfulness programs. *Mindfulness, 14*, 255–278. https://doi.org/10.1007/s12671-022-01997-2

Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology, 34*(1), 215–220. https://doi.org/10.1093/ije/dyh299

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*(2), 305–314. https://doi.org/10.1037/0033-2909.110.2.305

Braun, V., & Clarke, V. (2025). Reporting guidelines for qualitative research: A values-based approach. *Qualitative Research in Psychology, 22*(2), 399–438. https://doi.org/10.1080/14780887.2024.2382244

Browne, K. M., Drewes, A., Smalley, G., & Lichtenwalner, S. (2026). Data description, claim, evidence, reasoning (DCER): An instructional framework to develop undergraduates' data literacy skills and scientific reasoning. *Journal of College Science Teaching, 55*(3), 235–250. https://doi.org/10.1080/0047231X.2026.2633359

Chien, S.-C., & Li, W.-Y. (2024). Perceptions of supervisors and their doctoral students regarding the problems in writing the doctoral dissertation results section. *English for Specific Purposes, 76*, 14–27. https://doi.org/10.1016/j.esp.2024.06.003

Curtis, K., Chong, S.-W., & Kong, M. S. (2025). A qualitative synthesis of research into the use of exemplars in the English for Academic Purposes context to develop student feedback literacy. *Research Synthesis in Applied Linguistics, 1*(2), 302–339. https://doi.org/10.1080/29984475.2025.2495271

Diamantopoulos, A., & Winklhofer, H. M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research, 38*(2), 269–277. https://doi.org/10.1509/jmkr.38.2.269.18845

Du, H., & List, A. (2024). Evidence-based reasoning: Results from an intervention. *Applied Cognitive Psychology, 38*(5), Article e4238. https://doi.org/10.1002/acp.4238

Duncan, R. G., & Chinn, C. A. (2025). Evaluating the quality of argumentation: The role of epistemic ideals and reliable processes. *Cognition and Instruction, 43*(3), 201–232. https://doi.org/10.1080/07370008.2025.2497240

Flowerdew, L., & Petrić, B. (2024). A critical review of corpus-based pedagogic perspectives on thesis writing: Specificity revisited. *English for Specific Purposes, 76*, 1–13. https://doi.org/10.1016/j.esp.2024.05.003

Golparvar, S. E., Hu, G., & Seyedi, S. E. (2025). Cohesion in the discussion section of research articles: A cross-disciplinary investigation. *English for Specific Purposes, 77*, 1–19. https://doi.org/10.1016/j.esp.2024.08.004

Grainger, P., Scott, J. J., Mulgrew, K., Dean, M., Carey, M. D., & Johnston, C. (2025). Developing a formative assessment criteria-referenced tool (FACT) for doctoral students. *Educational Research and Evaluation, 30*(7–8), 748–765. https://doi.org/10.1080/13803611.2025.2556048

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, *P* values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology, 31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Grohnert, T., Gromotka, L., Gast, I., Delnoij, L., & Beausaert, S. (2024). Effective master's thesis supervision: A summative framework for research and practice. *Educational Research Review, 42*, Article 100589. https://doi.org/10.1016/j.edurev.2023.100589

Harsch, C., Koval, V., Kanistra, P. V., & Delgado-Osorio, X. (2024). Validating an integrated reading-into-writing scale with trained university students. *Assessing Writing, 62*, Article 100894. https://doi.org/10.1016/j.asw.2024.100894

Hu, G., & Bonsu, E. M. (2025). A cross-disciplinary study of value arguments in doctoral theses submitted to universities in Hong Kong. *English for Specific Purposes, 79*, 1–16. https://doi.org/10.1016/j.esp.2025.02.002

Jose, L. S. (2026). Improving methodological coherence in thesis proposal design: Development and preliminary validation of the Research Coherence Alignment Framework. *International Journal of Learning, Teaching and Educational Research, 25*(6), 588–615. https://doi.org/10.26803/ijlter.25.6.25

Jönsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

Jusslin, S., & Hilli, C. (2024). Supporting bachelor's and master's students' thesis writing: A rhizoanalysis of academic writing workshops in hybrid learning spaces. *Studies in Higher Education, 49*(4), 712–729. https://doi.org/10.1080/03075079.2023.2250809

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kurt, B., & Kafes, H. (2025). The role of disciplinary enculturation in stance-taking in L2 academic writing. *Acta Psychologica, 260*, Article 105720. https://doi.org/10.1016/j.actpsy.2025.105720

Levitt, H. M., Bamberg, M., Creswell, J. W., Frost, D. M., Josselson, R., & Suárez-Orozco, C. (2018). Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: The APA Publications and Communications Board task force report. *American Psychologist, 73*(1), 26–46. https://doi.org/10.1037/amp0000151

Lymer, G., Lindwall, O., & Greiffenhagen, C. (2024). Student writing in higher education: From texts to practices to textual practices. *Linguistics and Education, 80*, Article 101247. https://doi.org/10.1016/j.linged.2023.101247

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

McNeish, D., & Stapleton, L. M. (2016). Modeling clustered data with very few clusters. *Multivariate Behavioral Research, 51*(4), 495–518. https://doi.org/10.1080/00273171.2016.1167008

McShane, B. B., Bradlow, E. T., Lynch, J. G., Jr., & Meyer, R. J. (2024). "Statistical significance" and statistical reporting: Moving beyond binary. *Journal of Marketing, 88*(3), 1–19. https://doi.org/10.1177/00222429231216910

Musaeus, P., Prilop, C. N., Dikilitaş, K., Sundset, M. A., Svensen, C., Ershova, T., Boud, D., Lindberg, A. B., Bjælde, O. E., Gray, R. M., Jr., İstencioğlu, T., Kvernenes, M., Lassesen, B., & Raaheim, A. (2026). Teacher feedback literacy in higher education: Navigating enablers and constraints. *Assessment & Evaluation in Higher Education, 51*(2), 315–331. https://doi.org/10.1080/02602938.2025.2591556

Panadero, E., Delgado, P., Zamorano-Sande, D., Pinedo, L., Fernández Ortube, A., & Barrenetxea-Mínguez, L. (2025). Putting excellence first: How rubric performance level order and feedback type influence students' reading patterns and task performance. *Learning and Instruction, 99*, Article 102168. https://doi.org/10.1016/j.learninstruc.2025.102168

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

Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health, 30*(4), 459–467. https://doi.org/10.1002/nur.20199

Praestegaard Larsen, B. (2025). Research process model for bachelor's thesis. *Journal of Learning Development in Higher Education*, (34). https://doi.org/10.47408/jldhe.vi34.1379

Qiu, H., Cook, A. J., & Bobb, J. F. (2024). Evaluating tests for cluster-randomized trials with few clusters under generalized linear mixed models with covariate adjustment: A simulation study. *Statistics in Medicine, 43*(2), 201–215. https://doi.org/10.1002/sim.9950

Quinlan, K. M., & Pitt, E. (2025). Evaluative feedback isn't enough: Harnessing the power of consequential feedback in higher education. *Assessment & Evaluation in Higher Education, 50*(4), 577–591. https://doi.org/10.1080/02602938.2024.2430596

Rutkowski, D., Rutkowski, L., Thompson, G., & Canbolat, Y. (2024). The limits of inference: Reassessing causality in international assessments. *Large-scale Assessments in Education, 12*, Article 9. https://doi.org/10.1186/s40536-024-00197-9

Teng, M. F., & Ma, M. (2024). Assessing metacognition-based student feedback literacy for academic writing. *Assessing Writing, 59*, Article 100811. https://doi.org/10.1016/j.asw.2024.100811

Toulmin, S. E. (2003). *The uses of argument* (Updated ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511840005

Yan, D. (2024). Rubric co-creation to promote quality, interactivity and uptake of peer feedback. *Assessment & Evaluation in Higher Education, 49*(8), 1017–1034. https://doi.org/10.1080/02602938.2024.2333005