

# Diagnosing upper-secondary students' critical thinking in organic chemistry: a three-tier multiple-choice test on hydrocarbons

Muntholib Muntholib<sup>1,1</sup>, Hanifah Ariani Mahmudah<sup>1</sup>, Munzil Munzil<sup>1</sup>, M. Muchson<sup>1</sup>, Nur Candra Eka Setiawan<sup>1</sup>

<sup>1</sup>Chemistry Department, Faculty of Mathematics and Natural Science, Universitas Negeri Malang, Jl. Semarang No. 5, Malang 65145, Indonesia.

**Abstract.** This research examines upper-secondary students' critical thinking regarding hydrocarbons through a three-tier multiple-choice test (TT-MCT) that corresponds with six indicators: interpretation, analysis, evaluation, inference, explanation, and self-regulation. The instrument consisted of 20 five-option items with keyed reasons and a confidence tier. An expert review confirmed content validity ( $VR = 3.83$ ), and internal consistency during the main administration was high (Cronbach's  $\alpha = 0.911$ ). There were 85 Grade XI students from a public school who had finished the necessary prerequisite topics. Descriptive analyses revealed an overall mean of 59.55%, categorized as sufficient. Students did better on explanation (63.84%) and interpretation (63.12%) but worse on inference (53.88%) and analysis (54.35%). These findings indicate that learners more easily process descriptive and explanatory cues than those that require relational reasoning and the formulation of evidence-based conclusions. The three-tier format further revealed mixed answer–reason patterns (e.g., correct answers paired with flawed reasoning), indicating that many students can retrieve declarative knowledge but struggle to justify conclusions, particularly in mechanism-rich topics such as free-radical substitution. The findings endorse pedagogical designs that enhance the visibility and calibration of reasoning, including argumentation/claim–evidence–reasoning tasks focused on core mechanisms, targeted self-explanation prompts in worked examples, explicit coordination of multiple representations, and the formative application of confidence data to mitigate overconfidence and address misconceptions. The TT-MCT was helpful for assessing how well people were doing and for identifying actionable goals to improve analysis and inference in organic chemistry.

## 1 Introduction

Cultivating students' critical thinking is a central aim of chemistry education. We adopt Facione's six-indicator framework—interpretation, analysis, evaluation, inference, explanation, and self-regulation—as the operational definition for measurement and reporting in this study [1]. To capture not only what students answer but also why and how sure they are, we use a three-tier multiple-choice test (TT-MCT) consisting of Tier-1 content

\*Corresponding author: [muntholib\\_fmipa@um.ac.id](mailto:muntholib_fmipa@um.ac.id)

choices, Tier-2 reason choices, and Tier-3 confidence ratings. Three-tier instruments extend the two-tier diagnostic approach. They have demonstrated strong utility for profiling understanding and misconceptions across various science domains, including chemistry (e.g., states of matter) [2,3]. Confidence data are interpreted following the Certainty of Response Index (CRI) logic to differentiate low knowledge from robust misconceptions (i.e., high-confidence wrong) [4]. Recent reviews further document the growth and diagnostic value of multi-tier assessments in science education.

Hydrocarbons are a gateway to mechanistic reasoning in organic chemistry. Prior work shows students can often “get the product” yet struggle to justify steps mechanistically or coordinate evidence with representations (e.g., interpreting curved-arrow formalisms, connecting mechanisms to reaction-coordinate diagrams) [5]. Given the prevalence of radical chain processes in introductory organic chemistry, explicit attention to initiation–propagation–termination sequences and tracking of radical species is pedagogically important.

As a supporting theoretical lens, we draw on DeFT (Design–Functions–Tasks) for learning with multiple representations: well-designed tasks should help learners coordinate symbolic equations, particulate models, and mechanistic diagrams rather than treating any representation in isolation [4]. We also leverage evidence that self-explanation prompts (targeted, inference-oriented) yield reliable gains in understanding and transfer, which motivates the inclusion of explicit reason options and confidence ratings in the TT-MCT.

Aim and research questions. Building on these frameworks, this study (i) profiles students’ indicator-wise critical-thinking performance on hydrocarbons using a TT-MCT, and (ii) characterizes prominent answer–reason–confidence patterns to inform instruction on mechanism-rich topics. Concretely:

1. What are the indicator-wise performance levels (means/SDs) on the TT-MCT?
2. Which reasoning/confidence patterns are most prevalent, and what do they imply for instruction on radical substitution and related mechanisms?

The present study uniquely combines indicator-wise profiling with three-tier reasoning–confidence patterns. This dual approach contributes both diagnostically and pedagogically, offering a finer-grained lens into students’ critical thinking on mechanism-rich organic topics.

## **2 Method**

### **2.1 Research design**

We employed a descriptive, quantitative diagnostic design to profile indicator-wise critical thinking using a Three-Tier Multiple-Choice Test (TT-MCT). The three-tier format was chosen to elicit not only correctness but also reasoning (Tier-2) and confidence (Tier-3), allowing analysis of accuracy–justification–certainty patterns, consistent with contemporary test-development guidance and classical-test-theory (CTT) practice [9].

### **2.2 Participants and context**

The study involved Grade XI (upper-secondary) students from a public school who had covered the required prerequisites in chemistry. From the population, intact classes were selected, and a simple random sample was then drawn to produce the final analytic sample ( $N = 85$ ) [6]. Cluster-to-random workflows of this sort are standard in school-based surveys/sampling.

## 2.3 Instrument development

The TT-MCT comprised 20 five-option items (A–E) aligned to Facione’s six indicators (interpretation, analysis, evaluation, inference, explanation, and self-regulation). Each item supplied a keyed content option (Tier-1) and a keyed reason option (Tier-2). Tier-3 captured confidence on an ordered scale for later diagnostic use. Item writing and revision followed established test-development recommendations (content representativeness, clarity, plausible distractors, and alignment of reasons with mechanistic logic) [9]. While previous TT-MCTs have explored student misconceptions, this study extends the approach by integrating critical-thinking indicators and confidence ratings in the context of radical mechanisms in hydrocarbons. The contribution is thus both contextual and diagnostic.

### 2.3.1 Content validity

A panel of chemistry education experts (university and senior high school teachers) reviewed representativeness, clarity, terminology, and the mapping of reasons to the keyed concept. We summarized ratings using a content-validity ratio-style index and adopted the Lawshe decision logic, considering updated critical values [7,8] to judge item retention/revision before pilot testing.

### 2.3.2 Pilot testing and item analysis

A small pilot cohort (non-sample) completed the draft to examine empirical validity, reliability, difficulty ( $p$ ), discrimination (e.g., upper–lower 27%/point-biserial), and distractor functioning (non-functioning distractors <5% endorsement revised/removed). Retention criteria followed CTT conventions,  $p \approx 0.30$ – $0.70$  as acceptable difficulty and  $D \geq 0.30$  as good discrimination prior to the main administration [9,10].

## 2.4 Main administration and reliability

The final form was administered during regular lesson time in a single session (~70 minutes). Internal consistency was estimated with Cronbach’s  $\alpha$  on the study sample; interpretation followed current recommendations ( $\alpha$  as scale reliability, not item-quality proof, and the need to inspect item statistics alongside  $\alpha$ ).

## 2.5 Scoring and classification

Tier-1 and Tier-2 were scored using a rubric that rewarded accurate answers with correct reasons more strongly than partially correct patterns (e.g., correct answer paired with incorrect reason). Raw scores were summed and converted to percentages at the indicator and overall levels. Tier-3 confidence was not added to the numeric score; instead, it supported diagnostic interpretation (distinguishing low-knowledge vs high-confidence wrong). Where helpful, we report 95% confidence intervals for means ( $\text{mean} \pm 1.96 \cdot \text{SD} / \sqrt{n}$ ) to reflect estimation precision. Methodologically, using CTT item statistics (difficulty, discrimination, distractor analysis) in tandem with scale reliability aligns with standard practice and recent applied reports. For example, an item in which a student selects the correct content answer (Tier-1) and the correct reason (Tier-2) is awarded full credit. If the student selects the correct answer but an incorrect reason, partial credit is given. No credit is assigned if either tier is incorrect. This rubric ensures alignment with the assessment’s reasoning-based goals.

## 2.6 Data analysis

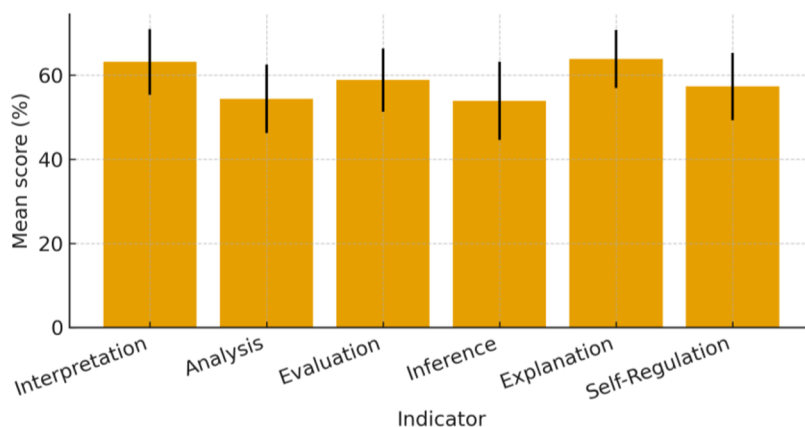
We computed descriptive statistics (means, SDs, CIs) for total and indicator scores; inspected item difficulty/discrimination and distractor efficiency; and summarized prominent three-tier patterns (answer–reason–confidence) to inform instruction. Analytical decisions (e.g., acceptable  $p$  and  $D$  thresholds, use of upper–lower 27%) were guided by recognized handbooks and applied medical-education item-analysis sources [9-18].

## 3 Result and Discussion

Table 1 below summarizes students' indicator-wise performance on the TT-MCT ( $N = 85$ ). The overall mean was 59.55% (classified as sufficient). The highest indicator mean occurred on explanation (63.84%), closely followed by interpretation (63.12%), while the lowest was in inference (53.88%) and analysis (54.35%). Figure 1 visualizes the indicator means with standard deviations.

**Table 1.** Descriptive statistics of critical-thinking indicators (TT-MCT;  $N = 85$ ).

Indicator	Mean (%)	SD	Category
Interpretation	63.12	7.80	Sufficient
Analysis	54.35	8.10	Low
Evaluation	58.82	7.50	Low
Inference	53.88	9.30	Low
Explanation	63.84	6.90	Sufficient
Self-regulation	57.29	8.00	Low



**Fig. 1.** Indicator-wise mean scores (%) on the TT-MCT. The bars represent the mean percentage scores for each critical-thinking indicator, and the whiskers indicate standard deviations, enabling comparison across indicators.

The comparatively lower scores on inference and analysis may reflect the cognitive demands associated with these indicators. Inference often requires conditional reasoning and the evaluation of multiple plausible alternatives, while analysis involves causal chaining, multi-variable integration, and recognition of underlying mechanisms. These tasks place a

higher cognitive load compared to more descriptive tasks like interpretation and explanation, potentially explaining the observed performance gap.

Our indicator-wise profile shows higher performance on explanation and interpretation, lower on analysis and inference, aligns with research showing that many learners can recall/descriptively explain outcomes in organic chemistry yet struggle with mechanistic integration (tracking causal relations, coordinating multiple factors) [12]. These difficulties are well documented for electron-pushing formalisms and multivariate mechanism problems, where students often treat arrows as surface cues rather than causal representations [12].

The three-tier format proved particularly useful for distinguishing mixed understanding (e.g., correct answer + flawed reason) from robust misconceptions, a distinction that matters because miscalibration (high confidence in wrong answers) is common and instruction-relevant. Recent evidence shows that positively biased confidence can persist across assessments and undermine learning, while targeted instruction can reduce overconfidence alongside improving performance [13-14].

### 3.1 Pedagogical implications

First, the pattern suggests a need to design for visible reasoning rather than answer-getting:

- Argument-Driven Inquiry (ADI). Integrating ADI cycles in chemistry labs promotes participation in scientific argumentation and improves the quality of student arguments; syntheses and classroom implementations (including secondary contexts) report gains in reasoning and critical-thinking outcomes [11]. For hydrocarbons, ADI tasks can center on competing mechanistic claims (e.g., radical selectivity, step order) with explicit warrants and evidence.
- Representational alignment. Activities should coordinate Johnstone's triplet (macro-submicro-symbolic) so that mechanism diagrams, particulate models, and equations are read jointly, not in isolation; this helps move students from descriptive recall to analytic inference [14].
- Calibration routines using confidence. Short cycles of prediction → confidence rating → feedback → reflection have been shown to improve monitoring accuracy and reduce overconfidence in STEM contexts; embedding such cycles around mechanism problems can convert Tier-3 data into formative feedback [13-14].

It is important to note that these implications are drawn from a single-site sample and should be interpreted as descriptive insights rather than generalized effects.

### 3.2 Domain-specific considerations

Within hydrocarbons, radical chain reactions require learners to maintain a coherent representation of initiation-propagation-termination and to track radical carriers across steps. Emphasizing causal links (what bonds break/form and why) and prompting students to justify each step can reduce superficial arrow-pushing and support movement on the analysis/inference indicators.

### 3.3 Limitations and next steps

Our inferences are bounded by a single-site, topic-bounded design. Future work should: (i) evaluate brief ADI units embedded in hydrocarbons with pre-post indicator-wise outcomes; (ii) include calibration interventions explicitly (confidence-guided reflection), and (iii) extend mechanism-focused tasks to other organic topics to test generality [11, 13-14].

### 3.4 Contribution to Chemistry Education Research

This study contributes to Chemistry Education Research (CER) by demonstrating how indicator-specific performance combined with multi-tier diagnostic tools can yield nuanced insights into students' reasoning processes. The integration of confidence metrics with content–reason pairings provides a richer diagnostic lens that is applicable to other challenging topics in organic chemistry and beyond.

## 4 Conclusion

Using a three-tier multiple-choice test (TT-MCT), this study profiled upper-secondary students' critical-thinking performance on hydrocarbons across six indicators. The indicator-wise pattern was consistent: students performed comparatively higher on explanation and interpretation, and lower on analysis and inference. Beyond mean differences, the three-tier format revealed mixed answer–reason patterns—evidence that many learners can retrieve descriptive knowledge yet struggle to justify conclusions and integrate mechanistic evidence. These findings validate TT-MCT as a classroom-friendly approach that estimates performance levels while exposing reasoning weaknesses that are otherwise hidden in single-tier scores. Pedagogically, the results support instruction that makes reasoning visible and calibrated: incorporating argument-and-evidence tasks around core organic mechanisms, embedding targeted self-explanation prompts in worked examples, and orchestrating multiple representations (symbolic, particulate, and mechanistic) to move learners from descriptive recall toward analytic inference. Using confidence data formatively can also surface overconfidence and misconception hotspots and guide feedback.

This work is bounded by a single-site, topic-focused design. Future research should examine multi-site implementations, track changes longitudinally after brief design interventions (e.g., ADI cycles complemented with self-explanation and calibration routines), and explore finer-grained diagnostic modeling to localize specific reasoning bottlenecks within analysis and inference. Overall, the study demonstrates that a TT-MCT-based approach can provide actionable evidence to strengthen critical-thinking outcomes in organic chemistry.

## Acknowledgment

**Acknowledgements:** We thank the Directorate General of Higher Education, Ministry of Education and Culture of the Republic of Indonesia for financial support this research. Directorate of Research and Community Service, Directorate General of Research and Development, Ministry of Higher Education, Science, and Technology Catalyst Research Scheme (Strategic Research Collaboration) Contract Number: 085/C3/DT.05.00/PL/2025, April 4th 2025.

## References

1. P. A. Facione, *Critical Thinking: What It Is and Why It Counts*, Insight Assessment. <https://www.law.uh.edu/blakely/advocacy-survey/Critical%20Thinking%20Skills.pdf>
2. I. Caleon and R. Subramaniam, “Development and Application of a Three-Tier Diagnostic Test to Assess Secondary Students' Understanding of Waves,” *Int. J. Sci. Educ.*, **32**(7), 939–961 (2010). <https://doi.org/10.1080/09500690902890130>

3. Z. D. Kirbulut, "Using Three-Tier Diagnostic Test to Assess Students' Misconceptions of States of Matter," *Eurasia J. Math. Sci. Technol. Educ.*, **10**(5), 509–521 (2014). <https://doi.org/10.12973/eurasia.2014.1128a>
4. S. Ainsworth, "DeFT: A Conceptual Framework for Learning with Multiple Representations," *Learn. Instr.*, **16**, 183–198 (2006). <https://doi.org/10.1016/j.learninstruc.2006.03.001>
5. M. Popova and S. L. Bretz, "Organic Chemistry Students' Challenges with Coherence Formation between Mechanisms and Reaction-Coordinate Diagrams," *Chem. Educ. Res. Pract.*, **19**, 732–740 (2018). <https://doi.org/10.1039/C8RP00064F>
6. S. L. Lohr, *Sampling: Design and Analysis* (2nd ed.), Chapman & Hall/CRC (2021).
7. C. H. Lawshe, "A Quantitative Approach to Content Validity," *Personnel Psychology*, **28**(4), 563–575 (1975). <https://doi.org/10.1111/j.1744-6570.1975.tb01393.x>
8. C. Ayre and A. J. Scally, "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 (2014). <https://doi.org/10.1177/0748175613513808>
9. S. M. Downing and T. M. Haladyna (eds.), *Handbook of Test Development*, Lawrence Erlbaum (2006).
10. N. I. Rashwan, S. R. Aref, O. A. Nayel, and M. H. Rizk, "Postexamination item analysis of undergraduate pediatric multiple-choice questions exam: implications for developing a validated question bank," *BMC Medical Education*, **24**, Article 168 (2024). <https://doi.org/10.1186/s12909-024-05153-3>
11. J. P. Walker and V. Sampson, "Argument-Driven Inquiry as a Way to Help Students Write to Learn by Learning to Write in General Chemistry Laboratory," *J. Res. Sci. Teach.*, **50**(5), 561–596 (2013). <https://doi.org/10.1002/tea.21082>
12. I. Caspari and N. Graulich, "Scaffolding the structure of organic chemistry students' multivariate comparative mechanistic reasoning," *Int. J. Phys. Chem. Educ.*, **11**(2), 29–42 (2019). <https://doi.org/10.12973/ijpce/211359>
13. I. Testa, A. Colantonio, S. Galano, I. Marzoli, F. Trani, & U.S. Uccio, "Effects of instruction on students' overconfidence in introductory quantum mechanics," *Physical Review Physics Education Research* (2020). <https://doi.org/10.1103/PhysRevPhysEducRes.16.010143>
14. B. Eilks and B. Byers (eds.), *Innovative Methods of Teaching and Learning Chemistry in Higher*, Royal Society of Chemistry, 2015.