IJE
Interdiciplinary Journal of Education
https://skillerindonesia.id/index.php/ije/index

Vol. 4, No 2: July 2026
E-ISSN: 2988-7798
pp. 162-172
Email:ije@skillerindonesia.com

Three-Dimensional Virus Models For Critical Thinking In Biology Education: An Indicator-Level Secondary Analysis

Baiq Fatmawati[1*], Fina Nawarul Uyun Fitriyah [2]
[1*,2] Faculty of Mathematics and Natural Sciences, Universitas Hamzanwadi, Indonesia

*Corresponding Author Email: baiq.fatmawati@hamzanwadi.ac.id

Abstract: This study presents an indicator-level secondary analysis of de-identified pretest-posttest data from 40 Grade 10 students (20 experimental and 20 control) who studied virus concepts at two madrasahs. The experimental group received instruction incorporating three-dimensional virus models, whereas the control group received lecture-based instruction. Four constructed-response items assessed interpretation, analysis, evaluation, and inference on a 1-4 scale. Raw gains and individual normalized gains were summarized descriptively. Because gain scores were non-normally distributed, two-sided Mann-Whitney U tests served as the primary between-group comparisons; Hedges' g with bootstrap 95% confidence intervals and pretest-adjusted ANCOVA were used as complementary analyses. The experimental group showed larger mean raw gains for interpretation (0.35 vs. 0.25), analysis (0.60 vs. 0.10), evaluation (0.35 vs. 0.20), inference (0.75 vs. 0.35), and the total score (2.05 vs. 0.90). Statistically significant between-group differences were observed for analysis (U = 271.0, p = .036; g = 0.74, 95% CI [0.16, 1.40]) and the total score (U = 320.5, p < .001; g = 1.30, 95% CI [0.78, 1.97]). Inference showed the largest descriptive improvement and a moderate effect (g = 0.56), although the between-group difference was not statistically significant (p = .136). ANCOVA yielded the same substantive pattern, with significant adjusted differences for analysis and the total score but not for interpretation, evaluation, or inference. Within the constraints of a small, non-randomized, two-school design, the findings suggest that model-supported instruction was most consistently associated with analytical reasoning, whereas interpretation and evaluation may require more explicit scaffolding, argumentation, and evidence-evaluation tasks.

Article History
Received: 10-05-2026
Revised: 11-06-2026
Published: 20-07-2026

Keywords:
3D virus models, biology education, critical thinking, model-based learning, secondary analysis.

How to Cite: Fatmawati, B., & Fitriyah, F. N. U. (2026). Three-Dimensional Virus Models For Critical Thinking In Biology Education: An Indicator-Level Secondary Analysis. *IJE : Interdisciplinary Journal of Education*, *4*(2), 162–172. https://doi.org/10.61277/ije.v4i2.317
https://doi.org/10.61277/ije.v4i2.317

1. Introduction

Critical thinking is a central objective of science and biology education because students must be able to interpret scientific information, analyze relationships among concepts, evaluate claims and evidence, and draw defensible conclusions. These processes are particularly important when learners encounter phenomena that are invisible or microscopic. Viral structure and replication cannot be observed directly in a typical school classroom; consequently, students must construct mental representations from diagrams, explanations, and models. When instruction relies predominantly on verbal exposition and two-dimensional images, learners may memorize labels such as capsid, nucleic acid, envelope, tail, and tail fibers without developing a coherent understanding of how these components relate to viral form and function.

Recent research continues to support active, discussion-rich approaches to science instruction. Lagubeau et al. (2020) found that active learning reduced academic risk among students whose formal scientific reasoning was still developing, whereas Ma et al. (2023) reported a positive association between students' critical thinking and physics achievement in a large school sample. Visual and verbal representations can further strengthen such engagement by making students' reasoning accessible for discussion and critique. Tytler et al. (2020) showed that drawing can support reasoning when learners construct and discuss explanatory representations; Park et al. (2021) identified a recurring plan-draw-evaluate pattern in collaborative science work; and Sjøberg et al. (2023) demonstrated that drawings, talk, and gestures jointly contribute to model-based reasoning in biology.

Three-dimensional models are therefore particularly relevant to the teaching of virus concepts. Recent studies in molecular, cellular, and secondary-school biology suggest that physical or digitally supported 3D representations can enhance visualization, engagement, communication, and understanding of structure-function relationships when students actively manipulate models and respond to guided questions (Howell et al., 2020; Wollmuth et al., 2023; Andic et al., 2024). Recent curriculum-oriented work further illustrates how 3D-printed biological models can be integrated with inquiry-based worksheets to make complex molecular structures more accessible to secondary students (Diao et al., 2026). However, improvement in an aggregate score does not reveal whether all dimensions of critical thinking respond equally to model-supported instruction.

Critical thinking is multidimensional and domain sensitive. Interpretation involves clarifying meaning and identifying relevant information; analysis involves recognizing relationships among components, claims, and evidence; evaluation requires judging the credibility and adequacy of reasons; and inference involves drawing warranted conclusions from the available evidence. Ma et al. (2023) emphasized critical thinking as an explicit outcome of science education, while Fauziah et al. (2024) concluded that critical thinking and scientific argumentation are interdependent and jointly support deeper science learning. A tangible virus model may directly facilitate analysis and inference by making structural relationships visible, whereas evaluation may require additional tasks involving competing explanations, argumentation, or systematic appraisal of evidence.

The present article reanalyzes the de-identified dataset from Fitriyah's (2025) undergraduate study, which examined the use of virus teaching aids in relation to students' critical-thinking performance. Whereas the original study emphasized the overall comparison between model-supported and conventional instruction, the present secondary analysis addresses a more diagnostic question: which critical-thinking indicators exhibited the strongest and weakest changes? The analysis additionally incorporates individual normalized-gain calculations, distribution-sensitive between-group tests, small-sample-corrected effect sizes with confidence intervals, and pretest-adjusted sensitivity analyses. Accordingly, this study aims to compare changes in interpretation, analysis, evaluation, and inference between the experimental and control groups while explicitly distinguishing descriptive improvement, effect magnitude, and statistical significance.

2. Literature Review

2.1 Visual Representation and Model-Based Reasoning in Biology

Biological structures often operate at scales that students cannot observe directly. Visual representations enable learners to examine spatial organization, relationships, and mechanisms, but their effectiveness depends on active meaning making. Tytler et al. (2020) found that student-generated drawings can function as reasoning tools when learners explain and negotiate the meaning of a representation. Park et al. (2021) further showed that collaborative visual work develops through repeated cycles of planning, drawing, and evaluation, while Sjøberg et al. (2023) documented how drawings, talk, and gestures complement one another during model-based reasoning in biology. Collectively, these studies indicate that visual representations are most productive when students are required to interpret, discuss, revise, and justify them.

Empirical research provides further support for the educational use of 3D representations. Howell et al. (2020) reported greater learning gains when students used interactive modules incorporating 3D-printed protein models to connect molecular structure with function. Wollmuth et al. (2023) found significant improvements in knowledge and confidence after students used 3D cellular models and engaged with structured small-group questions. In an inclusive biology classroom, Andic et al. (2024) reported that 3D modeling and printing supported description, visualization, misconception correction, and classroom communication, while cautioning that inaccurate scale or positioning may introduce new misconceptions. Taken together, these findings suggest that virus models can reduce representational difficulty, provided that their design is scientifically accurate and their classroom use is pedagogically guided.

2.2 Critical-Thinking Indicators and Instructional Alignment

Critical thinking in science is best conceptualized as a set of related but distinguishable cognitive processes. Interpretation and analysis enable students to make sense of information and representations, whereas evaluation requires judgments about the credibility and adequacy of evidence. Inference involves selecting relevant evidence and drawing warranted conclusions. Ma et al. (2023) showed that critical thinking is meaningfully associated with science achievement, and Fauziah et al. (2024) synthesized evidence indicating that critical thinking develops through close interaction with scientific argumentation, analytical questioning, discussion, and learning technologies. Indicator-level analysis is therefore valuable because a visual model may support some reasoning processes more directly than others.

Model-supported instruction should not be expected to generate critical thinking automatically. Research on scientific argumentation indicates that conceptual development is strengthened when students compare claims, use evidence, and justify their decisions through structured discourse (Li et al., 2022; Felton et al., 2022). Tang et al. (2022) likewise cautioned that students require explicit guidance regarding what constitutes evidence and how evidence should be used to support an argument. Thus, 3D virus models are most likely to promote analysis and inference when learners actively inspect, compare, and explain structural relationships. Evaluation, by contrast, requires supplementary prompts that ask students to determine whether a claim is accurate, assess whether the available evidence is sufficient, and justify why one explanation is stronger than another.

2.3 Research Gap and Analytical Contribution

Existing studies of virus teaching aids generally report learning outcomes or aggregate critical-thinking scores, providing limited insight into differential change across specific indicators. Moreover, raw gains can be difficult to interpret when pretest scores differ, and effect-size estimates may be unstable in small samples. The present secondary analysis addresses this gap by reporting indicator-level pretest and posttest scores, raw gain, individual normalized gain, nonparametric group comparisons, Hedges' g with bootstrap confidence intervals, and ANCOVA as a pretest-adjusted sensitivity analysis. Although this analytical triangulation cannot overcome the limitations of the original design, it provides a more transparent, nuanced, and appropriately cautious account of the available evidence.

3. Research Method

3.1 Design, Data Source, and Participants

This study conducted a secondary quantitative analysis of archival data from a quasi-experimental pretest-posttest control-group design. The dataset comprised de-identified scores from 40 Grade 10 students: 20 students from MA NWDI Wakan constituted the experimental group, and 20 students from MA NW Penendem constituted the control group. The groups had been selected purposively in the original study. Because participants were drawn from different schools and were not randomly assigned, contextual factors such as teacher characteristics, classroom climate, school resources, and prior learning may have confounded the observed differences. The present analysis therefore characterizes associations within the available sample rather than asserting unequivocal causality. Baseline comparisons revealed no statistically significant pretest differences between groups for interpretation, analysis, evaluation, inference, or total score (all $p > .05$). Nevertheless, the absence of statistically significant baseline differences does not establish group equivalence, particularly with only 20 students per group. Pretest-adjusted ANCOVA was therefore included as a sensitivity analysis, and potential school-level confounding was retained as a major study limitation.

3.2 Instructional Conditions

The experimental condition incorporated three-dimensional virus models as concrete visual supports, whereas the control condition relied on conventional lecture-based instruction. The archived manuscript and model photographs indicate that the models represented several viral forms and components, including capsids, genetic material, tails, tail fibers, envelopes, and surface structures. During model-supported instruction, the intended learning activities included observing structures, identifying components, comparing viral forms, discussing structure-function relationships, and drawing conclusions from visual evidence. These activities align most directly with the critical-thinking processes of interpretation, analysis, and inference.

Table 1. Documented and Unavailable Characteristics of the Original Intervention

Aspect	Experimental condition	Control condition / archival limitation
Learning resource	Three-dimensional physical virus models representing major forms and components.	No physical model; lecture-based explanation.
Documented cognitive activities	Observation, identification, comparison, discussion of structural relationships, and conclusion drawing.	Listening, note-taking, and teacher explanation are described; the exact sequence was not archived.
Intervention dose	Not sufficiently documented in the secondary dataset.	Not sufficiently documented in the secondary dataset.
Teacher equivalence	Not verifiable from the available dataset.	Not verifiable from the available dataset.
Worksheets / prompts	The archived article indicates that guided questions and discussion were recommended supports, but the exact worksheet and prompt set was not retained.	No comparable archived prompt set was available.

Because the archival dataset did not preserve the number or duration of instructional sessions, the identity of the instructor(s), a fidelity checklist, or the complete set of worksheets, the intervention cannot be fully replicated from the available records. Rather than filling these gaps with assumptions, this limitation is reported explicitly. Future studies should employ a standardized lesson plan, equivalent instructional time across conditions, the same teacher where feasible, and a fidelity log documenting how each critical-thinking indicator is addressed during instruction.

3.3 Critical-Thinking Instrument

The archival test comprised four constructed-response items, with one item representing each critical-thinking indicator: interpretation, analysis, evaluation, and inference. Each response was scored using a four-point rubric (1-4), yielding a maximum score of 4 for each indicator and 16 for the total score. The indicators and

IJE
Interdiciplinary Journal of Education
https://skillerindonesia.id/index.php/ije/index
Vol. 4, No 2: July 2026
E-ISSN: 2988-7798
pp. 162-172
Email:ije@skillerindonesia.com

scoring descriptors followed the operational framework documented in the original thesis. Constructed-response items are appropriate for eliciting students' explanations of relationships and justifications of conclusions; however, the use of only one item per indicator restricts score reliability and content coverage. The available dataset did not include the original expert-validation forms, item-revision history, or independent double-scoring records. Consequently, content-validity coefficients and inter-rater reliability could not be reconstructed. As a limited check of internal consistency across the four items, Cronbach's alpha was 0.69 at pretest and 0.71 at posttest for the combined sample. These estimates should be regarded as preliminary because each indicator was represented by a single item and the indicators are conceptually distinct. Indicator-level reliability cannot be estimated from a single item; therefore, the resulting scores warrant cautious interpretation. The original thesis further documents that the four-item instrument underwent theoretical and construct review by expert raters as well as empirical pilot testing. Item validity was examined using Pearson product-moment correlations, reliability using Cronbach's alpha, and item quality using difficulty and discrimination indices. The instrument blueprint and four-point analytic scoring rubric were also documented. However, the available thesis file does not report the identities of the raters or the numerical item-level validity, reliability, difficulty, and discrimination coefficients. These values therefore cannot be independently verified, and the incomplete psychometric reporting is acknowledged as a measurement limitation.

3.4 Data Analysis

For each student, raw gain was calculated by subtracting the pretest score from the posttest score. Individual normalized gain was calculated as $g_i = (post_i - pre_i)/(\text{maximum score} - pre_i)$, using a maximum score of 4 for each indicator and 16 for the total score. The reported group N-gain represents the mean of the individual N-gain values rather than an N-gain calculated from group means. This distinction explains why the control group's mean analysis score increased from 1.50 to 1.60 while its mean individual N-gain was -0.02: several students with relatively high pretest scores declined, and individual normalization assigned proportionally greater weight to those declines. Shapiro-Wilk tests indicated non-normal gain distributions for all indicator scores in both groups ($p < .05$). Accordingly, two-sided Mann-Whitney U tests were used as the primary between-group comparisons. Independent-samples t tests were reported only as robustness checks because Levene's tests indicated homogeneity of variance for the four indicators but not for total gain. Effect magnitude was summarized using Hedges' g, a small-sample-corrected standardized mean difference, with percentile-bootstrap 95% confidence intervals based on 10,000 resamples. Finally, ANCOVA models compared posttest scores between groups while controlling for the corresponding pretest score. Because group assignment was confounded with school, ANCOVA was interpreted as a sensitivity analysis rather than as evidence of causality. Statistical significance was evaluated at $\alpha = .05$.

3.5 Ethical Considerations

The secondary analysis used a de-identified spreadsheet and involved no new contact with students. Names were excluded from the analytical dataset used to generate the results reported in this article. Because the participants were Grade 10 students and may have included minors, school permission, student assent, parental consent, and any relevant institutional ethics approval from the original study should be reported explicitly wherever documentation is available. The original study was conducted at MA NWDI Wakan and MA NW Penendem during the first semester of the 2025/2026 academic year following school-level administrative authorization. The present secondary analysis used only de-identified student scores and involved no additional interaction with participants. The available thesis does not provide an institutional ethics-committee approval number or describe written parental-consent and student-assent procedures in sufficient detail. Accordingly, no formal ethics approval number is reported, and the incomplete documentation of consent procedures is acknowledged as a study limitation.

IJE
Interdiciplinary Journal of Education
https://skillerindonesia.id/index.php/ije/index
Vol. 4, No 2: July 2026
E-ISSN: 2988-7798
pp. 162-172
Email:ije@skillerindonesia.com

4. Results and Discussion

4.1 Results

4.1.1 Three-Dimensional Virus Models

The physical models used in the experimental class represented multiple viral forms and made normally microscopic components visible to students. The models provided a shared object for observation and discussion, enabling students to identify structural features and compare their spatial organization.

Figure 1. Three-dimensional virus models used in the experimental class

4.1.2 Descriptive Change and Baseline Profile

Baseline comparisons revealed no statistically significant differences between the experimental and control groups for any critical-thinking indicator or for the total pretest score (all $p > .05$), although the experimental group had a slightly higher mean total score than the control group (6.45 vs. 5.95). Following the intervention, the experimental group demonstrated larger mean raw gains across all four indicators and in the total score. The largest descriptive gain in the experimental group occurred for inference (0.75), followed by analysis (0.60), whereas interpretation and evaluation each increased by 0.35. In the control group, mean raw gains ranged from 0.10 for analysis to 0.35 for inference. Complete descriptive statistics are presented in Table 2.

Table 2. Pretest, Posttest, Raw Gain, and Mean Individual N-Gain

Indicator	Exp. pre M (SD)	Exp. post M (SD)	Exp. gain M (SD)	Exp. N-gain	Ctrl. pre M (SD)	Ctrl. post M (SD)	Ctrl. gain M (SD)	Ctrl. N-gain
Interpretation	1.70 (0.66)	2.05 (0.76)	0.35 (0.67)	0.12	1.40 (0.50)	1.65 (0.67)	0.25 (0.64)	0.08
Analysis	1.60 (0.50)	2.20 (0.70)	0.60 (0.60)	0.25	1.50 (0.61)	1.60 (0.60)	0.10 (0.72)	-0.02
Evaluation	1.65 (0.59)	2.00 (0.86)	0.35 (0.67)	0.13	1.60 (0.60)	1.80 (0.70)	0.20 (0.52)	0.07
Inference	1.50 (0.51)	2.25 (0.97)	0.75 (0.79)	0.32	1.45 (0.60)	1.80 (0.62)	0.35 (0.59)	0.12
Total score	6.45 (1.61)	8.50 (2.44)	2.05 (1.05)	0.23	5.95 (1.70)	6.85 (1.73)	0.90 (0.64)	0.09

To facilitate visual comparison of students' improvement, Figure 2 presents the mean raw gain for each critical-thinking indicator and the total score in the experimental and control groups. Descriptively, the experimental group exhibited higher mean gains across all indicators. Inference showed the largest gain in the experimental group, followed by analysis, while the largest overall between-group difference was observed for the total score.

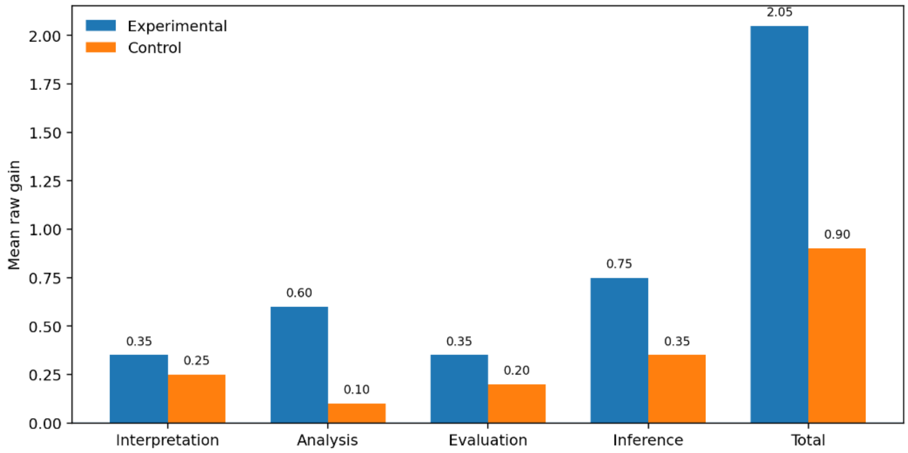

Figure 2. Comparison of Mean Raw Gains Between the Experimental and Control Groups

4.1.3 Between-Group Tests, Effect Sizes, and ANCOVA

The primary Mann-Whitney U analysis identified statistically significant between-group differences in gain scores for analysis (U = 271.0, p = .036) and the total score (U = 320.5, p < .001). No statistically significant between-group differences were observed for interpretation or evaluation. Inference showed the largest indicator-level descriptive gain difference (0.40) and a moderate standardized effect (g = 0.56); however, the Mann-Whitney U test did not indicate a statistically significant difference (U = 250.5, p = .136). The independent-samples t test produced the same substantive conclusion: the gain difference for analysis was statistically significant (p = .022), whereas the difference for inference did not reach the .05 significance threshold (p = .076). Complete results are reported in Table 3.

Table 3. Between-Group Comparisons of Gain Scores

Indicator	Mann–Whitney U	p	Hedges' g	Bootstrap 95% CI	ANCOVA group p
Interpretation	210.0	.770	0.15	[-0.49, 0.76]	.304
Analysis	271.0	.036	0.74	[0.16, 1.40]	.006
Evaluation	221.5	.499	0.24	[-0.37, 0.88]	.415
Inference	250.5	.136	0.56	[0.00, 1.19]	.068
Total score	**320.5**	**< .001**	**1.30**	**[0.78, 1.97]**	**< .001**

The ANCOVA sensitivity analysis was consistent with the gain-score findings. After adjustment for the corresponding pretest score, the group effect was statistically significant for analysis (F(1, 37) = 8.40, p = .006) and the total posttest score (F(1, 37) = 15.97, p < .001). The adjusted group effect for inference approached but did not reach statistical significance (F(1, 37) = 3.52, p = .068), whereas interpretation and evaluation remained nonsignificant. Taken together, these analyses provide the strongest evidence for improvement in analytical reasoning and the overall critical-thinking score. The inference result should therefore be interpreted as a promising descriptive pattern rather than as a statistically confirmed effect.

4.2 Discussion

4.2.1 Why Analysis Showed the Most Consistent Evidence

Analysis was the only individual indicator for which the between-group difference was supported by both the nonparametric gain comparison and the pretest-adjusted ANCOVA. This pattern is theoretically plausible because the 3D models externalized the component parts and spatial relationships of viruses. Students could decompose each model into constituent structures, compare viral forms, and relate capsids, genetic material, envelopes, tails, and surface structures to their respective functions. Such activities align directly with analytical reasoning, which requires identifying components and relationships rather than merely recalling labels. This interpretation is consistent with recent research on model-based learning. Howell et al. (2020) found that interactive 3D models helped students translate molecular structures into functional explanations, while Wollmuth et al. (2023) reported significant learning gains when students manipulated scaled cellular models and responded to guided questions. Andic et al. (2024) similarly showed that 3D modeling can support visualization and communication in biology, and Sjøberg et al. (2023) demonstrated that model-based reasoning develops through coordinated visual, verbal, and gestural activity. In the present study, the virus models likely functioned as cognitive anchors for observation, comparison, and explanation. Nevertheless, the available data do not permit the effect of the models to be isolated from peer discussion, teacher questioning, novelty, or other active-learning processes.

4.2.2 Inference: Large Descriptive Gain but Statistical Uncertainty

Inference exhibited the largest descriptive increase in the experimental group and a moderate Hedges' g, suggesting that students may have benefited from drawing conclusions from visible structural evidence. A model can help learners move from an observed feature - for example, the presence of an envelope or tail fibers - to a tentative explanation concerning attachment, host interaction, or viral form. This interpretation is consistent with research indicating that model-based reasoning advances when learners coordinate visual representations with exploratory talk and evaluative processes (Park et al., 2021; Sjøberg et al., 2023). However, the between-group difference for inference did not reach statistical significance in the Mann-Whitney test ($p = .136$), the t test ($p = .076$), or ANCOVA ($p = .068$). The result should therefore not be interpreted as conclusive evidence that the model caused an improvement in inference. The combination of a moderate effect estimate, a wide confidence interval, and a small sample indicates substantial uncertainty. A larger and better-controlled study is needed to determine whether the observed pattern is robust or partly attributable to sampling variation, school context, or teacher effects.

4.2.3 Why Interpretation and Evaluation Improved Less

Interpretation improved only modestly. Although a model can help students recognize component parts, translating a visual form into accurate biological meaning requires explicit representational guidance. Students need to understand what color, shape, scale, and spatial placement signify and to connect those visual cues with appropriate disciplinary terminology. Tytler et al. (2020) and Park et al. (2021) showed that visual representations become effective reasoning tools when learners explain, evaluate, and revise them through discourse. Andic et al. (2024) further cautioned that inaccurate scale, positioning, or model quality may generate new misconceptions. Worksheets that require students to label components, paraphrase meanings, examine scale, and explain the role of each structure may therefore strengthen interpretation. Evaluation likewise showed only a small, nonsignificant effect. This pattern is theoretically plausible because evaluation involves judging the credibility of claims, comparing competing explanations, identifying limitations, and justifying decisions with evidence. A physical model provides an object for observation, but it does not automatically generate competing claims or criteria for judgment. Recent reviews indicate that critical thinking and scientific argumentation are mutually reinforcing

(Fauziah et al., 2024), while classroom studies emphasize responsive dialogue and explicit attention to evidence (Felton et al., 2022; Tang et al., 2022). Socratic questioning may also encourage learners to examine alternatives and reflect on the quality of their reasoning (Favero et al., 2025). Future model-supported lessons should therefore incorporate misinformation cases, alternative virus diagrams, evidence-ranking tasks, and written claim-evidence-reasoning prompts.

4.2.4 Alternative Explanations and Pedagogical Implications

The observed differences cannot be attributed solely to the physical models. The groups were drawn from different schools, and the archival records do not establish whether the same teacher, instructional duration, curriculum pacing, or questioning strategy was used across conditions. Increased attention, novelty, peer interaction, and students' prior familiarity with virus content may also account for part of the observed pattern. The findings therefore support an association between the model-supported instructional package and stronger analytical outcomes within this sample, rather than an isolated causal effect of the models themselves. In practice, virus models should be embedded within a structured reasoning sequence rather than used merely as visual displays. One appropriate sequence would require students to observe and label components, compare two viral forms, explain structure-function relationships, evaluate competing statements, and formulate a conclusion supported by specific evidence from the models. Park et al. (2021) showed that planning, visual construction, and evaluation can organize the development of ideas, while Li et al. (2022) and Felton et al. (2022) emphasized that argumentation and responsive dialogue require purposeful prompts and teacher guidance. Such a progression integrates representation, explanation, evidence appraisal, and reflection.

4.3 Study Limitations

Several limitations constrain the interpretation of these findings. First, the sample was small (n = 40), resulting in wide effect-size confidence intervals and limited statistical power. Second, the experimental and control groups were drawn from different schools, meaning that school context and treatment condition were confounded. Third, the archival dataset did not retain a detailed intervention log, information on teacher equivalence, instructional duration, or fidelity measures. Fourth, each critical-thinking indicator was represented by only one item, precluding indicator-level reliability analysis; moreover, the original content-validity and inter-rater records were unavailable. Fifth, the study assessed only immediate pretest-posttest change and did not examine retention or transfer to novel biological contexts. Sixth, mean individual N-gain may behave unexpectedly when students begin with different pretest scores or exhibit score declines; it should therefore be interpreted alongside raw gain and model-based analyses. Finally, the findings are context specific and should not be generalized beyond comparable Grade 10 biology settings without replication.

5. Conclusion

Within this small, non-randomized, two-school sample, instruction supported by three-dimensional virus models was associated with larger gains than lecture-based instruction across all measured critical-thinking indicators. The most consistent evidence was observed for analysis: the experimental group demonstrated a larger gain, a moderate Hedges' g, a statistically significant Mann-Whitney comparison, and a significant pretest-adjusted group effect. Inference showed the largest descriptive improvement and a moderate effect estimate, but the between-group difference did not reach conventional statistical significance. Interpretation and evaluation improved only modestly and did not differ significantly between groups. These findings suggest that 3D virus models are particularly well suited to activities requiring students to identify components and relate structure to function, whereas interpretation and evaluation require more explicit representational prompts, argumentation, and evidence-judgment tasks. Stronger causal conclusions will require larger samples, same-school or randomized comparisons, equivalent teachers and instructional time, validated multi-item measures for each indicator, documented intervention fidelity, and delayed posttests.

6. Recommendation

Biology teachers are encouraged to integrate three-dimensional virus models with structured worksheets, evidence-based questioning, collaborative discussion, and reflective explanation so that the models function as tools for reasoning rather than as visual displays alone. Future studies should recruit larger and more comparable samples, document instructional duration and intervention fidelity, employ instruments with verified validity and inter-rater reliability, and include delayed assessments to determine whether improvements in analysis and inference are sustained over time.

Data availability

The de-identified dataset used in this secondary analysis is available from the corresponding author upon reasonable request, subject to permission from the participating schools and applicable privacy requirements.

Declaration of the use of artificial intelligence (AI)

Generative AI tools were used to assist with English-language editing, statistical analysis, and graph preparation. The authors reviewed and verified all AI-assisted outputs and remain fully responsible for the accuracy, interpretation, and originality of the manuscript.

Funding

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

Acknowledgment

The authors gratefully acknowledge the participating schools, biology teachers, and students who contributed to the original study.

References

Andic, B., Lavicza, Z., Ulbrich, E., Cvjeticanin, S., Petrovic, F., & Maricic, M. (2024). Contribution of 3D modelling and printing to learning in primary schools: A case study with visually impaired students from an inclusive biology classroom. *Journal of Biological Education*, 58(*4*), 795-811. https://doi.org/10.1080/00219266.2022.2118352.

Diao, K., Duff, S. M. G., & Li, H. (2026). An idea to explore: Enhancing the teaching of genome editing through 3D printed models of CRISPR-Cas9 technology. *Biochemistry and Molecular Biology Education*, 54(*3*), 292-296. https://doi.org/10.1002/bmb.70047.

Fauziah, A. N. M., Wasis, Hendratmoko, A. F., Mahdiannur, M. A., Ermawan, M. Z. F., Suwandi, E., & Ratri, S. Y. (2024). Relationship between critical thinking and scientific argumentation in science learning. *Jurnal Pendidikan IPA Indonesia*, 13(2), 239-254. https://doi.org/10.15294/jpii.v13i2.2585.

Favero, L., Pérez-Ortiz, J.A., Käser, T., Oliver, N. (2025). Enhancing Critical Thinking in Education by Means of a Socratic Chatbot. In: Bellas, F., Fontenla-Romero, O. (eds) AI in Education and Educational Research. AIEER 2024. Communications in Computer and Information Science, vol 2519. Springer, Cham. https://doi.org/10.1007/978-3-031-93409-4_2.

Felton, M., Levin, D. M., De La Paz, S., & Butler, C. (2022). Scientific argumentation and responsive teaching: Using dialog to teach science in three middle-school classrooms. *Science Education*, 106(*6*), 1354-1374. https://doi.org/10.1002/sce.21740.

Fitriyah, F. N. U. (2025). Pengaruh penggunaan alat peraga virus terhadap kemampuan berpikir kritis siswa [Undergraduate thesis, Universitas Hamzanwadi]. Repository Universitas Hamzanwadi.

Howell, M. E., Booth, C. S., Sikich, S. M., Helikar, T., van Dijk, K., Roston, R. L., & Couch, B. A. (2020). Interactive learning modules with 3D printed models improve student understanding of protein structure-function relationships. *Biochemistry and Molecular Biology Education*, 48(*4*), 356-368. https://doi.org/10.1002/bmb.21362.

Lagubeau, G., Tecpan, S., & Hernandez, C. (2020). Active learning reduces academic risk of students with non-formal reasoning skills: Evidence from an introductory physics massive course in a Chilean public university. *Physical Review Physics Education Research*, 16, 023101. https://doi.org/10.1103/PhysRevPhysEducRes.16.023101.

Li, X., Wang, W., & Li, Y. (2022). Systematically reviewing the potential of scientific argumentation to promote multidimensional conceptual change in science education. *International Journal of Science Education*, 44(*7*), 1165-1185. https://doi.org/10.1080/09500693.2022.2070787.

Ma, X., Zhang, Y., & Luo, X. (2023). Students' and teachers' critical thinking in science education: Are they related to each other and with physics achievement?. *Research in Science & Technological Education*, 41(2), 734-758. https://doi.org/10.1080/02635143.2021.1944078.

Park, J., Tang, K.-S., & Chang, J. (2021). Plan-Draw-Evaluate (PDE) pattern in students' collaborative drawing: Interaction between visual and verbal modes of representation. *Science Education*, 105(5), 1013-1045. https://doi.org/10.1002/sce.21668.

Sjøberg, M., Furberg, A., & Knain, E. (2023). Undergraduate biology students' model-based reasoning in the laboratory: Exploring the role of drawings, talk, and gestures. *Science Education*, 107(*1*), 124–148. https://doi.org/10.1002/sce.21765

Tang, X., Levin, D. M., Chumbley, A. K., & Elby, A. (2022). Arguing about argument and evidence: Disagreements and ambiguities in science education research and practice. *Science Education*, 106(*2*), 285-311. https://doi.org/10.1002/sce.21696.

Tytler, R., Prain, V., Aranda, G., Ferguson, J., & Gorur, R. (2020). Drawing to reason and learn in science. *Journal of Research in Science Teaching*, 57(*2*), 209-231. https://doi.org/10.1002/tea.21590.

Wollmuth, E. M., Correa, A., Alvarado Obando, M., Smith, M. K., Buckley, D. H., Hefferon, K. L., & Angert, E. R. (2023). Helping students see bacteria in 3D: Cellular models increase student learning about cell size and diffusion. *Journal of Microbiology & Biology Education*, 24(*3*), e00089-23. https://doi.org/10.1128/jmbe.00089-23.