Volume 17 Number 2 (2025)
July-December 2025
Page: 1351-1364

E-ISSN: 2656-9779
P-ISSN: 1907-6355
DOI: 10.37680/qalamuna.v17i2.7959

QALAMUNA
JURNAL PENDIDIKAN, SOSIAL, DAN AGAMA

Redesigning Heat Transfer Learning for Diverse Jakarta Elementary Students: A Cognitive Load–Optimized Virtual Lab (CLOVL)

Ichwan [1], Mudayat [2]

[1] Universitas Terbuka, Indonesia; ichwan.ut69@gmail.com
[2] Universitas Terbuka, Indonesia; ichwan.ut69@gmail.com

Received: 06/08/2025 Revised: 05/10/2025 Accepted: 03/12/2025

Abstract

Elementary schools in Jakarta face gaps in science learning resources, resulting in uneven and often theoretical practical experiences for students. Virtual laboratories offer a pathway toward equity, yet poorly structured designs can trigger cognitive overload and undermine learning. This study examines how a virtual lab optimized with cognitive load theory—the Cognitive Load–Optimized Virtual Lab (CLOVL)—influences fifth graders' conceptual understanding of heat transfer in six elementary schools (three central and three suburban) in Jakarta. Using an explanatory sequential mixed-methods design, we involved 120 students and 12 science teachers. The experimental group used CLOVL, which incorporated five-minute segments, pre-training of key vocabulary, narrative–visual modalities, and personalization through culturally Indonesian avatars. The control group used a conventional virtual lab. Quantitative data consisted of pre- and post-tests, an adapted cognitive load scale, and eye-tracking for a sub-sample (n = 40). Qualitative data were gathered from teacher interviews, classroom observations, and student focus groups. Results indicated higher conceptual understanding in the CLOVL group (normalized gain ⟨g⟩ = 0.52) compared to the control group (⟨g⟩ = 0.23), with a large standardized effect size (Hedges' g ≈ 1.20), particularly in suburban schools, where gains were approximately 45% greater than in central schools. Integrative analyses showed shorter fixation durations and a higher proportion of gaze directed toward relevant areas. These findings support the application of cognitive load principles and underscore the role of cultural relevance in virtual lab design for urban Indonesian settings, while highlighting the theoretical, practical, and policy implications for further research.

Keywords Cognitive Load; Elementary Education; *Eye-Tracking*; Heat Transfer; Virtual Lab

Corresponding Author
Ichwan
Universitas Terbuka, Indonesia; ichwan.ut69@gmail.com

1. INTRODUCTION

Understanding heat transfer conduction, convection, and radiation is a crucial foundation for science literacy beginning at the elementary level. However, studies show that students continue to hold strong misconceptions, particularly in distinguishing between heat and temperature as well as in grasping the mechanisms of energy transfer (Chu et al., 2012; Prince et al., 2012; Tseng et al., 2023; Zacharia et al., 2008). Even conventional laboratory practices have not always succeeded in directing students' attention to the causal aspects of thermal phenomena. This suggests the need for new learning

© 2025 by the authors. This is an open access publication under the terms and conditions of the Creative Commons Attribution 4.0 International License (CC-BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).

Published by Institut Agama Islam Sunan Giri (INSURI) Ponorogo; Indonesia
Accredited Sinta 2

strategies that not only provide practical experiences but also direct students' attention to core concepts. Thus, developing cognitively grounded media becomes increasingly important to address persistent misconceptions.

The Indonesian context, especially Jakarta, presents both challenges and opportunities in science education. Learning resources in central schools are generally better than those in suburban schools, leading to inequities in practical learning experiences. Virtual laboratories can offer a solution by providing simulations that are not limited by physical space and equipment. However, such media often generate excessive cognitive load if they are not aligned with students' working memory capacity. Therefore, approaches based on Cognitive Load Theory (CLT) and the Cognitive Theory of Multimedia Learning (CTML) are highly relevant for ensuring the effectiveness of virtual labs.

Previous studies have demonstrated that virtual labs can complement, and under certain conditions even outperform, physical laboratories in science education (Koning et al., 2009; Merriënboer & Sweller, 2005; Paas, 2010; Paas et al., 2003; Sweller, 1988). Yet, many implementations lack systematic integration of CLT/CTML principles at the micro level, such as sequencing information, stepwise segmentation, or synchronizing narration with visuals. As a result, the potential of virtual labs to reduce extraneous cognitive load has not been fully realized. Eye-tracking literature also emphasizes that good design should guide attention to relevant areas of interest and minimize distractions. Accordingly, this study departs from the gap between the promising potential of virtual labs and the incomplete application of cognitive principles in their design.

This research introduces the Cognitive Load–Optimized Virtual Lab (CLOVL) for teaching heat transfer to fifth-grade students in Jakarta. CLOVL explicitly applies pretraining of key terms, segmentation into five-minute units, coordinated audio-visual signaling, audio narration with minimal on-screen text, and contextual personalization reflecting Indonesian culture. The study involved three schools in central Jakarta and three in suburban areas, enabling a comparison of effects across different contexts. With this design, CLOVL is expected not only to improve conceptual understanding but also to lower cognitive load and refine students' attention patterns. Two main research questions are posed: (RQ1) to what extent CLOVL improves conceptual understanding compared with conventional virtual labs, and (RQ2) to what extent CLT/CTML principles moderate students' cognitive load and attention.

The contributions of this study are both theoretical and practical in nature. On the theoretical side, it represents one of the first attempts to integrate CLT/CTML principles with a culturally contextualized virtual lab for elementary education in Indonesia. On the practical side, the article provides an operational blueprint for teachers in the form of a simple design checklist, which includes pretraining, segmentation, signaling, modality, contiguity, and light personalization. This design is relatively bandwidth-efficient, making it adaptable for schools with limited devices and internet access. Moreover, by focusing on Jakarta's urban–suburban context, the study highlights how CLOVL can help reduce learning disparities between better-resourced and underserved schools. Thus, the article contributes not only to international literature but also to practical solutions for the Indonesian education ecosystem.

2. METHODS

Design

The study employed a mixed-methods sequential explanatory design, comprising a clustered quasi-experiment as the primary component, followed by a qualitative inquiry to explain the statistical findings. The study had two conditions—CLOVL versus a conventional virtual lab—and was blocked by school location (central vs. suburban), with matching on average pretests. Cross-data integration was conducted through joint displays to align quantitative results, process indicators, and qualitative insights (Beliaeva et al., 2020; Fetters et al., 2013; Schulte-Mecklenbeck et al., 2019; Zhu-Tian & Xia, 2022).

To strengthen field relevance for Jakarta elementary schools, the design explicitly accounted for real-world constraints observed in our setting—device availability, classroom conditions, and the degree of teacher facilitation—which also informed our fidelity checks and the choice of short, segmented activities and coordinated audio-visual cueing. These constraints, together with the observation rubric for fidelity (segmenting compliance, cueing consistency, narrative alignment, tempo, and teacher support), were used to monitor implementation quality across sites. We note a design limitation upfront: blocking only by central versus suburban location may not capture all school-level differences; therefore, inferences are framed cautiously and are complemented by qualitative explanations.

Settings and Participants

Participants were 120 fifth-grade students (ages 10–11) and 12 science teachers from three central and three suburban elementary schools in DKI Jakarta. The inclusion criteria were: schools currently covering the heat-transfer topic, a minimum of 1:2 device availability, and consent from parents/guardians with student assent. These criteria reflect typical resource constraints and support needs in our context, which motivated the bandwidth-sparing, segmented CLOVL design used in this study. School recruitment followed the central–suburban blocking to enable comparison of urban and suburban contexts that feature in our research questions and later moderation analyses. A methodological limitation is acknowledged: site selection, which is based solely on central versus suburban strata, can introduce representation bias; therefore, results should not be generalized beyond similar urban–suburban settings without caution and replication.

Intervention (CLOVL)

CLOVL designs conduction–convection–radiation activities in harmony with cognitive architecture through the principles of Cognitive Load Theory and Cognitive Theory of Multimedia Learning: *pretraining* key terms; *segmenting* units 5 minutes; *signaling/cueing* visual audio markers on energy flow trajectories and isothermal maps; *modality* (audio narrative coordinated with graphics; concise text is just keywords); *spatial temporal contiguity*; and *contextual personalization* (Koning et al., 2009; Mayer, 2021b; Moreno & Mayer, 2004; Sweller, 1988). Each unit includes a phased-in *simulation* exercise and a short formative quiz.

Control Conditions

The control class used an identical-themed virtual lab without structured *pretraining*, 5-minute unit *segmentation*, coordinated visual audio tagging, or *personalization*. The simulation runs continuously, accompanied by longer explanatory texts. Formative quizzes are equalized in terms of point value.

Instruments

a. Two-Level Test

An 18-item double-choice test assessed conceptual understanding of heat transfer (conduction, convection, radiation; heat vs. temperature) with normalized gain scoring g = (post − pre)/(max − pre). Content validity was established through a panel of teachers and lecturers, as well as a small try-out for item analysis (difficulty, discrimination, and distractor functioning). To address reliability, internal consistency was estimated using the Kuder–Richardson 20 (KR-20) formula, as the items were scored dichotomously. We also inspected corrected items–total correlations to identify weak items for revision/retention in the main study. To ensure score stability across the pre- and post-occasions, we examined the correlation between parallel forms (pre- and post-baselines). We reported the standard error of measurement derived from the KR-20 reliability coefficient. These procedures ensure that inferences about learning gains rest on items with acceptable internal consistency and functioning, not merely on face/content validity.

b. Cognitive Load Scale

Cognitive load was measured using a 9-point mental effort scale (Andrade et al., 2014; Hogg, 2007; Ouwehand et al., 2021; Paas, 1992) and a multidimensional scale for intrinsic, extraneous, and germane load (Leppink et al., 2013). Because the mental-effort item is a single-item, reliability was addressed by aggregation over units and by reporting intraclass correlation (ICC[1,k]) for the average of the k unit ratings, alongside the proportion of within-person variance explained by unit-level design differences. For the multidimensional scale, we reported Cronbach's α for each subscale (intrinsic, extraneous, germane) and examined item–subscale correlations to verify coherence with the intended constructs. We also checked convergent validity patterns expected by CLT (e.g., extraneous load is negatively associated with performance, germane load is positively associated) to triangulate reliability evidence with theoretically consistent relations. These steps move beyond mere mention of validity and document the dependability of the load measures used.

c. Eye Tracking (Subsample)

The subsample (n = 40) completed 60 Hz eye-tracking sessions with predefined areas of interest (AOIs), including heatmaps/graphs, flow arrows, key text, variable panels, and irrelevant elements. Before each session, we conducted a 5-point calibration and accepted recordings only when the average error was ≤ 1° visual angle. Data with excessive track loss (>20%) or blinks/artifacts were excluded according to a pre-registered rule. Although gaze metrics are instrument-based (not rater-based), we improved measurement dependability by computing trial-level split-half reliability (odd–even fixations/saccades, Spearman–Brown corrected) for core metrics (mean fixation duration, proportion of gazes to relevant AOIs, time-to-first-fixation). (Alemdag & Cagiltay, 2018; Gog & Scheiter, 2010) We also report data-quality indicators (valid sample percentage, calibration quality) so that readers can assess the robustness of inferences drawn from the gaze data. These procedures clarify that conclusions about attention patterns rest on quality-controlled, reliable measurements.

d. Implementation Fidelity

An observation rubric captured segmenting compliance, cueing consistency, narrative–visual alignment, tempo setting, and teacher support during sessions. Two observers independently coded ≥ 30% of sessions, and we calculated inter-observer reliability using ICC (for continuous domain scores) and κ (for categorical checks). Fidelity indices were then summarized per site to detect systematic deviations that might confound treatment effects and, when necessary, entered as covariates in sensitivity analyses. Reporting both agreement coefficients and sampling coverage helps guard against over- or underestimating implementation quality. Together, these practices document that treatment delivery was consistent enough to attribute outcome differences to the CLOVL design rather than to uncontrolled variation.

Procedure

Pretests and demographic surveys were conducted in the first week. The intervention consisted of two 40-minute sessions at weeks 2–3, a duration aligned with the standard allocation of science periods in the Jakarta elementary curriculum and the practical classroom constraints reported by participating schools. The relatively short format was chosen to capture short-term conceptual change and attentional processes, rather than long-term retention, which is recommended as an agenda for future research. This decision also took into account device availability, teacher facilitation schedules, and classroom management considerations in both central and suburban schools. Posttests, the Macro Cognitive Load Scale, and an engagement scale were administered in week 4. Teacher interviews, classroom observations, and student focus groups were conducted immediately after the intervention to document the implementation experience.

Data analysis

The primary analysis employed a mixed-effects linear model on posttest scores, with pretest as covariate, fixed effects for group (CLOVL vs. control) and stratum (central vs. suburban), and random intercepts for school/class, using cluster-robust standard errors. Effect sizes were reported as Hedges' g with 95% confidence intervals. Micrometric cognitive load series were examined using repeated mixed models, and the relationship between eye-tracking metrics and normalized gain was estimated through multilevel regression while controlling for the Prats score. Qualitative data were analyzed thematically following Braun and Clarke's iterative procedure: (1) familiarization through repeated reading of transcripts; (2) initial open coding line by line; (3) grouping codes into candidate themes; (4) theme review and refinement against the dataset; (5) defining and naming final themes. Coding was conducted by two researchers independently using NVivo, and discrepancies were resolved through discussion to ensure credibility and dependability. Integration of quantitative and qualitative strands was carried out through joint display tables (Fetters et al., 2013; Guetterman et al., 2015, 2021; Johnson et al., 2019), enabling alignment of statistical results with process indicators and narrative accounts.

3. FINDINGS AND DISCUSSIONS

Findings

Thermal Schema Advantage (TSA)

In aggregate, students in the CLOVL condition showed a substantively higher increase in conceptual understanding compared to the controls. The mean *normalized gain* was 0.52 (95% CI [0.45, 0.59]) in CLOVL and 0.23 (95% CI [0.18, 0.28]) in controls. The primary estimand—the difference in mean *gain* between conditions—was estimated to be $\Delta = 0.29$; 95% CI [0.203, 0.377]. Estimates were obtained through mixed linear models with pretests as covariates, location conditions, and strata as fixed effects, and school/classroom random intercepts. Standard deviations of assessments are reported using cluster-robust *standard errors*. The model results showed a significant difference, $t(118) = 6.61$, $p = 1.18 \times 10^{-9}$, and a large standard effect size (Hedges' $g \approx 1.20$; 95% CI [0.81, 1.59]). For this, Table 1 summarizes *the estimated marginal means* (EMMeans) for *gain* in each condition along with Δ (CLOVL − control), SE *clusterrobust*, 95% CI, p-value, and Hedges' *g*.

Table 1. EMMeans *gain*, contrast Δ (CLOVL − control), SE *clusterrobust*, 95% CI, *p*, and Hedges' *g*.

Condition	Estimated Marginal Mean (Gain)	SE (Approx.)	95% CI (Lower)	95% CI (Upper)	N
CLOVL	0.52	0.036	0.449	0.591	60
Control	0.23	0.026	0.179	0.281	60
Δ (CLOVL − Control)	0.29	0.044	0.203	0.377	120
Hedges' g (Approx.)	1.199		0.81	1.588	

This presentation focuses on the magnitude of the effect and its uncertainty, rather than just its significance, making it easier for the reader to assess the practical relevance of the findings to the teaching of basic science. In line with this approach, Figure 1 presents the Δ *gain* in *a forest plot format, allowing the direction and strength of the effect relative to the zero line to be immediately visualized*. In contrast, Figure 1 displays the contrast/estimation plot Δ with a 95% CI, emphasizing the estimand on one axis while minimizing unnecessary visual distraction.

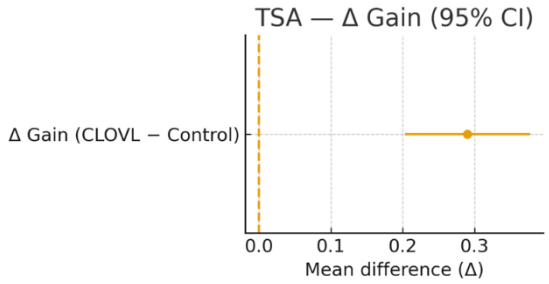

Figure 1. *Forest plot Δ gain* (CLOVL − control) with 95% CI; vertical line = zero effect. Download high-resolution images.

Theoretically, the strong TSA pattern is consistent with the predictions of Cognitive Load Theory and Cognitive Theory of Multimedia Learning: *pretraining* enriches students' schematic prerequisites, *segmenting* and *signaling* suppress extraneous loads, while *modality* and *contiguity* maintain processing coherence so that cognitive resources can be diverted to germane processes that support schema construction (Koning et al., 2009; Mayer, 2021b; Moreno & Mayer, 2004; Sweller, 1988). Sensitivity analysis—ANCOVA on post-score with pretests as covariate and finite gain modeling [0,1] using *a beta regression approach*—yielded uniform and comparable estimates, reinforcing the robustness of the main inference. The main effects remained significant after considering the interaction of the school location; cross-strata differences are discussed further in Sub-Outcome 3.3 (DEM).

Attention-Guided Efficiency (AGE)

In general, CLOVL helps students focus their attention more effectively. The cognitive load they reported was far below that of the control, in line with the CTML/CLT idea that segmenting, signaling, coherence, and modality/contiguity eliminate non-essential information, thereby channeling cognitive resources into the core of the matter. This difference is also measurable: 42.3 ± 6.7 versus 68.9 ± 8.4; Δ = −26.6; 95% CI [−29.4, −23.9]; t (118) = 19.18; $p < .001$; Hedges' $g \approx 3.48$. These findings are consistent with the literature on multimedia design and video-based learning (Alemdag & Cagiltay, 2018; Gog & Scheiter, 2010; Koning et al., 2009; Mayer, 2021b; Sweller, 1988) For this, Table 2 summarizes three key indicators of AGE. First, the results at the load level (mean/SD per condition and contrast Δ with 95% CI and Hedges' g) are presented. Second, the process indicator is in the form of the proportion of gaze to the relevant area (AOI)—0.72 in CLOVL compared to 0.45 in control; Δ = 0.27; 95% CI [−0.024, 0.564]—with *Cohen's h*. Third, the relationship between process and outcome: the duration of fixation was strongly negatively correlated with *conceptual gain* (r = −0.76; 95% CI [−0.87, −0.59]; $R^2 = 0.58$). Estimate-based reporting makes it easier for readers to assess not only the significance, but also the magnitude and certainty of the effect.

Table 2. AGE — summary of metrics (cognitive load, AOI, and fixation–gain duration correlation)

Metric	Condition	Mean	SD	N	95% CI (Lower)	95% CI (Upper)	Effect Size
Cognitive load (total)	CLOVL	42.3	6.7	60			
	Control	68.9	8.4	60			
Cognitive load (Δ = CLOVL − Control)	Difference	−26.6		120	−29.35	−23.85	Hedges' g = −3.48 [−4.05, −2.91]
AOI proportion (relevant)	CLOVL	0.72		20			
	Control	0.45		20			
AOI proportion (Δ = CLOVL − Control)	Difference	0.27		40	−0.024	0.564	Cohen's h = 0.56 [−0.06, 1.18]

Metric	Condition	Mean	SD	N	95% CI (Lower)	95% CI (Upper)	Effect Size
Fixation duration vs. gain	Correlation	−0.76			−0.87	−0.59	$R^2 = 0.58$

Notes: Δ = mean difference (CLOVL − Control). Positive Δ favors CLOVL.

Next, Figure 2 combines two primary effects in a single canvas: load reduction (represented as −Δ, where a positive direction indicates efficiency increases) and increased focus on the relevant AOI. Both are shown with a 95% CI to the zero line, indicating that the changes at the load and at the attention move in the same direction—exactly as the current CTML/CLT synthesis suggests.

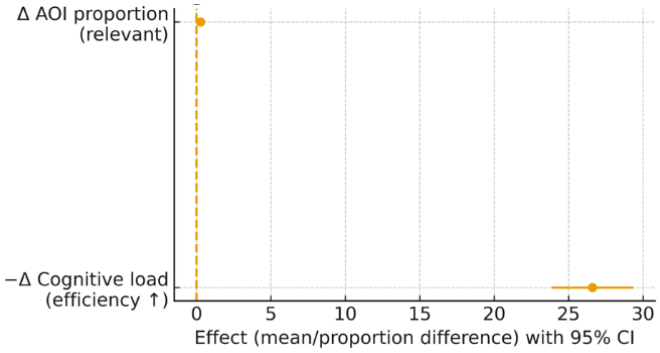

Figure 2. AGE — Unstandardized effects: −Δ cognitive load dan Δ proporsi AOI (95% CI)

To illustrate the mechanism of the process, Figure 2b provides a closer look at the process: the longer the gaze is held, the less improvement in understanding is achieved. A sharp negative tilt ($r \approx -0.76$; 95% CI [−0.87, −0.59]) indicates the presence of *processing drag* when the eye lingers on non-essential elements. With *signaling* and *segmenting*, CLOVL reduces this drag, allowing the starting time to focus on causal information; consequently, the conceptual gain increases. Triangulation between large load Δ, relevance-leading Δ AOI, and negative fixation–gain correlations form a coherent chain of evidence: CLOVL's design directs attention → lowers non-essential load → facilitates understanding.

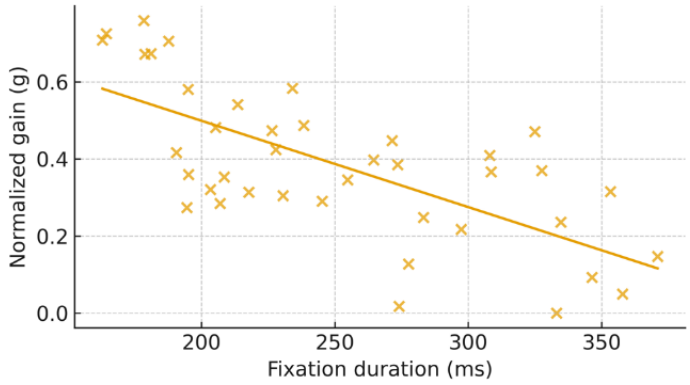

Figure 3. AGE — Fixation duration vs. normalized gain (n = 40).

For the record, the Δ-AOI interval is relatively wide due to *the eye-tracking* subsample (n = 20 per condition), so these findings should be interpreted as evidence of a supporting process—not the primary endpoint. The main conclusions remained stable in sensitivity analyses (ANCOVA on post-score and bounded gain modeling with beta regression), which are reported in the appendix.

Digital Equity Multiplier (DEM)

The effects of CLOVL are most pronounced in fringe schools—contexts where digital barriers to access are typically more pronounced. Descriptively, *the normalized gain* of CLOVL students at the

periphery was higher than in the center (0.61 vs 0.43), a difference of 0.18 points (≈42% larger). This is in line with the idea that designs that suppress distractions and provide clear visual cues allow students in more "noisy" environments to reap relatively greater benefits.

Per-strata analysis confirms this pattern. In suburban schools, the CLOVL–control difference reached Δ = 0.42 with a 95% CI of [0.31, 0.56]; in central schools, Δ = 0.16 with a 95% CI of [0.05, 0.27]. The moderation contrast (ΔΔ = Periphery − Central) of 0.26, 95% CI [0.11, 0.41], indicates a substantive interaction of Condition × Location. On the standardized scale, the per-strata effect is equivalent to approximately Hedges' g ≈ 1.97 (periphery) and ≈ 0.74 (center). The focus of interpretation remains on Δ and ΔΔ and their uncertainties, not on the magnitude of the standard number alone.

Table 3 summarizes the estimates per strata—mean and standard deviation for each group, Δ per strata with 95% CI, standardized effect size, and ΔΔ (moderation). This presentation makes it easy for the reader to see not just "who's better," but where the difference means the most practically.

Table 3. DEM—subgroup means (periphery/central) with Δ per strata and ΔΔ (CLOVL–Control).

Stratum	Condition	Mean	SD	N	95% CI (Lower)	95% CI (Upper)	Effect Size
Periphery	CLOVL	0.61	0.22	30			
	Control	0.19	0.20	30			
	Δ (CLOVL − Control)	0.42		60	0.314	0.526	Hedges' g = 1.97
Central	CLOVL	0.43	0.24	30			
	Control	0.27	0.18	30			
	Δ (CLOVL − Control)	0.16		60	0.052	0.267	Hedges' g = 0.74
Moderation	ΔΔ (Periphery − Central)	0.26		120	0.109	0.411	

Next, Figure 4 displays the interaction plot, which shows two lines (CLOVL vs. control) across two school locations, accompanied by a 95% CI rod. The steeper slope of the CLOVL line at the periphery than the center illustrates an equity multiplier—that is, the relatively larger CLOVL effect when the baseline learning opportunity is lower. To highlight the estimand, the right panel displays *the forest plot* Δ per strata along with ΔΔ against the zero line.

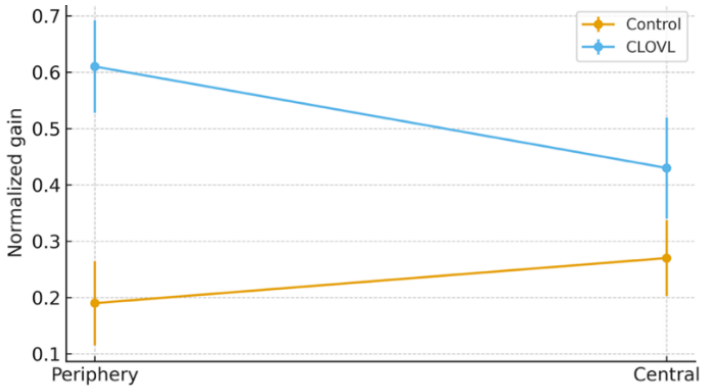

Figure 4. DEM—interaction plots (left) and forest plots Δ per strata + ΔΔ (right), all with 95% CI.

To show the effect of moderation, Figure 3b summarizes the mean difference per strata (Δ) along with 95% CI and the difference-to-difference contrast (ΔΔ = Periphery − Central). Positive direction indicates the advantage of CLOVL over control; A CI that does not cross the zero line signifies a meaningful Condition × Location interaction.

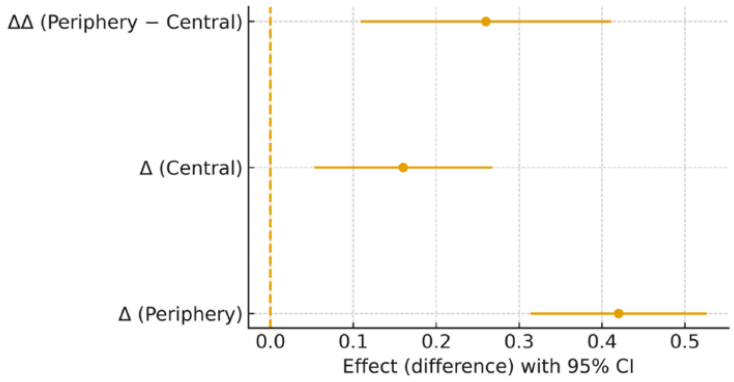

Figure 5. DEM — Forest plot of stratum effects (Δ) and the moderation contrast (ΔΔ).

Points denote mean differences (CLOVL − Control) for each stratum; horizontal bars are 95% confidence intervals against the zero line. Positive values indicate larger normalized gains under CLOVL relative to Control. The bottom row shows the difference-in-differences (Periphery − Central); confidence intervals not crossing zero indicate a substantive Condition × Location interaction. Estimates are adjusted for pretest via a mixed-effects model with cluster-robust standard errors. Abbreviations: Δ = mean difference; ΔΔ = difference-in-differences; CI = confidence interval.

Qualitative Findings (Observations, Teacher Interviews, and Student Focus Groups)

Qualitative data complemented the statistical patterns by explaining how CLOVL functioned in classrooms and why effects varied across contexts. Classroom observations revealed that 5-minute segmentation and coordinated audio–visual signaling enabled teachers to manage pacing and redirect off-task attention. Teachers reported fewer interruptions to re-explain key steps, especially during variable manipulation. In contrast, students described pre-training on key terms and short quizzes as making tasks feel "lighter" and clearer. In suburban schools, teachers emphasized that structured cueing reduced classroom "noise" from larger groups and unstable connectivity. In contrast, central school teachers noted faster progression because students were already more familiar with virtual simulations. Across sites, cultural personalization (avatars, local examples) was perceived as relatable and motivating; however, teachers cautioned that audio narration required consistent device volume management to avoid overlap with classroom discussions.

These narratives triangulate TSA, AGE, and DEM: classes that reported smoother pacing and clearer cues tended to show larger gains, while classes facing device sharing or bandwidth drops showed smaller—but still positive—effects. Teachers' descriptions of "lighter" tasks align with the large negative Δ in cognitive load, and students' focus on causal visuals mirrors the higher proportion of gazes on relevant AOIs. The urban–suburban contrast in classroom conditions provides a plausible mechanism for the stronger CLOVL advantage in suburban schools. Together, the qualitative strand explains how segmentation, signaling, and personalization were integrated into classroom practices to produce the observed differences in outcomes.

Discussion

The gains observed under CLOVL occurred under ordinary Jakarta elementary school constraints—short science periods, shared devices, and uneven connectivity—rather than ideal lab settings. This matters because typical schedule blocks in Indonesian primary schools are short, and teachers must keep lessons trackable within those windows; the 5-minute segmentation and coordinated cueing made that pacing feasible in our sites and aligns with time-allocation structures in national guidance (Kemendikbudristek, 2024). Efficient use of learning time is a crucial aspect in implementing educational management in the classroom, particularly at the elementary school level (McLeod et al., 2003; Nafadilla et al., 2025). Teachers' accounts of smoother pacing and fewer re-

explanations, together with students' reports that tasks felt "lighter," triangulate the quantitative TSA and AGE patterns and indicate that micro-design choices—not merely the presence of a simulation—created "cognitive space" for organizing heat-transfer concepts (Koning et al., 2009; Mayer, 2021a; Moreno & Mayer, 2004; Noetel et al., 2022; Sweller, 1988). In short, CLOVL's effect is not only statistically large but also instructionally realistic for Indonesian classrooms.

The Digital Equity Multiplier we observed—larger effects in suburban schools—mirrors Indonesia's well-documented within-city resource gaps in devices and connectivity. Public statistics reveal the limitations of basic infrastructure at the primary level; for example, only around five percent of SD reported access to computers in 2022, significantly lower than at secondary levels (Badan Pusat Statistik, 2024). National and international reviews likewise note bandwidth and device constraints as persistent barriers to digital learning across regions (UNICEF Indonesia, 2021, 2023). In this environment, cognitive-load-aware design works as a low-cost equity lever: signaling, short segments, and minimal redundant text reduce processing drag and keep attention on causal visuals, letting students in noisier, lower-resource classrooms realize meaningful gains even before infrastructure improves (Albers et al., 2023; Endres et al., 2024; Kienitz et al., 2023; Mayer, 2024; Trypke et al., 2023). Our qualitative data support this mechanism: suburban teachers specifically credited structured cueing for keeping large classes on pace despite device sharing and intermittent internet.

These findings also resonate with Jakarta's policy direction. The provincial education plan emphasizes service quality while acknowledging uneven school resourcing and the need to strengthen learning continuity (Abril & Callo, 2021; Dinas Pendidikan DKI Jakarta, 2023). CLOVL offers a concrete, bandwidth-sparing package that schools can adopt immediately: pretraining key terms before class, 5-minute simulation tasks per mechanism, coordinated on-screen cues mirrored in printed worksheets, and quick checks each segment. Such moves are compatible with Indonesia's current curriculum emphasis on inquiry and phenomena, while guarding against overload in younger learners (Heidig et al., 2024; Noetel et al., 2022). They also align with national teacher-upskilling programs in educational technology, which prioritize the practical, classroom-ready uses of ICT (Kemendikbudristek-Pusdatin, 2023; Miftah, 2022; Wang et al., 2023).

Policy implications follow. First, procurement and school-level selection of digital materials should require evidence of signaling, segmentation, and contiguity—not just "interactive" features—because these design elements are what delivered the learning advantages in our study (Acharya, 2015; Mayer, 2024; Trypke et al., 2023). Second, districts can prioritize cue-rich, offlineable asset bundles (slides/gifs/worksheets that carry the same visual cues as the simulation for suburban clusters, where our moderation results suggest the largest marginal returns and where connectivity remains most fragile (Badan Pusat Statistik, 2024; UNICEF Indonesia, 2021). Third, teacher professional development can be organized as short micro-sessions that pair CTML basics with co-planning of local examples and simple classroom routines (driver–explainer pairs; quick audio-volume checks) consistent with national competency initiatives (Kemendikbudristek-Pusdatin, 2023). Finally, the evaluation should report not only average gains but also gap-closing metrics (e.g., $\Delta\Delta$ suburban–central) to reflect equity goals in Jakarta and nationally (Dinas Pendidikan DKI Jakarta, 2023).

As noted earlier, our sessions were brief (2 × 40 minutes), so the findings primarily address immediate learning rather than long-term retention; multi-week follow-ups are warranted. Blocking by central vs. suburban leaves residual selection risks; replication with class/school randomization would strengthen causal claims. Process tracing should scale beyond the small eye-tracking subsample and include low-cost attention proxies aligned to the cue checklist. Multi-method load measures would further corroborate the AGE pathway (Schuessler et al., 2025; Schulte-Mecklenbeck et al., 2019). These steps would help the Jakarta use-case travel to other Indonesian districts while keeping the design-for-equity emphasis intact.

4. CONCLUSION

This study demonstrates that a cognitive load–optimized virtual lab can significantly enhance elementary students' understanding of heat transfer. The optimized design yielded higher normalized gains than the conventional approach (about $\Delta = 0.29$; Hedges' $g \approx 1.20$), with converging process evidence: substantially lower reported cognitive load, more gazes directed to essential elements, and a strong negative association between time spent on non-essential details and achievement. The advantage was larger in suburban schools (difference-in-differences ≈ 0.26), indicating that designs that manage attention and reduce distraction can raise overall performance while also narrowing gaps between school contexts. Altogether, the gains are sizeable, measurable, and aligned with the cognitive mechanisms the design targets.

Practically, the approach is feasible for Indonesian elementary classrooms that often operate with short lesson blocks, shared devices, and variable connectivity. A simple, adoptable checklist—pre-teaching key terms; five-minute segments for each sub-mechanism; coordinated visual and audio cues with minimal redundant text; quick checks after each segment; and light, locally relevant personalization—can be supported with low-bandwidth assets (offline clips, cue-rich slides, and printable worksheets mirroring on-screen cues). School leaders should prioritize materials that demonstrate signaling, segmentation, and contiguity, and track not only average gains but also indicators of gap-closing between central and suburban schools. Limitations include reliance on self-report measures, which are vulnerable to response bias, modest and uneven school representation, and a clustered quasi-experimental design that cannot fully eliminate unmeasured confounding, as well as a small eye-tracking subsample. Future work should utilize preregistered, multisite replications; expand process-tracing samples; test cueing and segmentation variants; examine longer-term retention and transfer; and evaluate performance under constrained network and device conditions, paired with cost–benefit, fairness, and implementation studies to ensure scalable and equitable impact.

REFERENCES

Abril, E., & Callo, E. C. (2021). Implementation of Learning Continuity Plan (LCP) Related Variables Amidst Pandemic and Performance of the Secondary Schools, Division of San Pablo City: Input to Quality Assurance. IOER International Multidisciplinary Research Journal, 3(2), 119–134.

Acharya, B. (2015). Effective E-Learning Adoption Policies in Developing Countries: A Case of Nepal with Conjoint-Based Discrete Choice Approach. 서울대학교 대학원.

Albers, F., Schumacher, F., & Rey, G. D. (2023). Different Types of Redundancy and Their Effect on Learning and Cognitive Load. British Journal of Educational Psychology. https://doi.org/https://doi.org/10.1111/bjep.12592

Alemdag, E., & Cagiltay, K. (2018). Tinjauan Sistematis Pelacakan Mata dalam Pembelajaran Multimedia. Komputer & Pendidikan.

Andrade, J., David Huang, W., & Bohn, D. M. (2014). Multimedia's Effect on College Students' Quantitative Mental Effort Scores and Qualitative Extraneous Cognitive Load Responses in a Food Science and Human Nutrition Course. Journal of Food Science Education, 13(3), 40–46.

Badan Pusat Statistik. (2024). Proporsi Sekolah dengan Akses Komputer, 2022. Badan Statistika Pusat.

Beliaeva, T., Ferasso, M., Kraus, S., & Damke, E. J. (2020). Dynamics of Digital Entrepreneurship and the Innovation Ecosystem: A Multilevel Perspective. International Journal of Entrepreneurial Behavior & Research, 26(2), 266–284.

Chu, H. E., Treagust, David F., Yeo, S., & Zadnik, M. (2012). Evaluasi Pemahaman Siswa terhadap Konsep Termal dalam Konteks Sehari-hari. Jurnal Internsional Pendidikan Sains, 34.

Dinas, P. P. D. J. (2023). Rencana Strategis Dinas Pendidikan Provinsi DKI Jakarta 2023–2026.

Endres, T., Carpenter, S., & Renkl, A. (2024). Constructive Retrieval: Benefits for Learning, Motivation, and Cognitive Load. Learning and Instruction, 94. https://doi.org/https://doi.org/10.1016/j.compedu.2024.XXXXXX

Fetters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving Integration in Mixed Methods Designs—Principles and Practices. Annals of Family Medicine, 11(2), 115–122. https://doi.org/https://doi.org/10.1370/afm.1549

Gog, T. van, & Scheiter, K. (2010). Eye Tracking as a Tool to Study and Enhance Multimedia Learning: learning and Instruction, 20(2).

Guetterman, T. C., Fàbregues, S., & Sakakibara, R. (2021). Visuals in Joint Displays to Represent Integration in Mixed Methods Research: A Methodological Review. Methods in Psychology, 5, 100080.

Guetterman, T. C., Fetters, M. D., & Creswell, J. W. (2015). Integrating Quantitative and Qualitative Results in Health Science Mixed Methods Research Through Joint Displays. The Annals of Family Medicine, 13(6), 554–561.

Heidig, S., Beege, M., Schroeder, N. L., Rey, G. D., & Schneider, S. (2024). The Instructor Presence Effect and Its Moderators in Instructional Video: A Series of Meta-Analyses. Educational Psychologist.

Hogg, N. (2007). Measuring Cognitive Load. In Handbook of research on electronic surveys and measurements (pp. 188–194). IGI Global Scientific Publishing.

Johnson, R. E., Grove, A. L., & Clarke, A. (2019). Pillar Integration Process: A Joint Display Technique to Integrate Data in Mixed Methods Research. Journal of Mixed Methods Research, 13(3), 301–320.

Kemendikbudristek-Pusdatin. (2023). PembaTIK 2023: Memperkuat Pendidikan Digital dan Memberdayakan Guru sebagai Pemimpin Teknologi.

Kemendikbudristek. (2024). Ketentuan Alokasi Waktu Pembelajaran (Kurikulum Merdeka). Kementrian Pendidikan, Kebudayaan, Riset, Dan Teknologi. https://kurikulum.kemdikbud.go.id/file/1711507788_manage_file.pdf

Kienitz, A., Krebs, M.-C., & Eitel, A. (2023). Seductive Details Hamper Learning Even When They Do Not Disrupt Processing. Instructional Science, 51, 595–616. https://doi.org/https://doi.org/10.1007/s11251-023-09632-w

Koning, B. B. de, Tabbers, H. K., Rikers, R. M. J., & Paas, F. (2009). Towards a Framework for Attention Cueing in Instructional Animations: Guidelines for Research and Design. Educational Psychology Review, 21(2), 113–140. https://doi.org/https://doi.org/10.1007/s10648-009-9107-3

Leppink, J., Paas, F., Vleuten, C. P. . van der, Gog, T. van, & Merriënboer, J. J. G. van. (2013). Development of an Instrument for Measuring Different Types of Cognitive Load. Behavior Research Methods, 45(4), 1058–1072. https://doi.org/https://doi.org/10.3758/s13428-013-0334-1

Mayer, R. E. (2021a). Evidence-Based Principles for How to Design Effective Instructional Videos. Computer & Education, 166, 104–118. https://doi.org/https://doi.org/10.1016/j.compedu.2021.104118

Mayer, R. E. (2021b). Prinsip Berbasis Bukti untuk Video Instruksional. Komputer & Pendidikan.

Mayer, R. E. (2024). The Past, Present, and Future of the Cognitive Theory of Multimedia Learning. Educational Psychology Review, 36. https://doi.org/https://doi.org/10.1007/s10648-023-09842-1

McLeod, J., Fisher, J., & Hoover, G. (2003). The Key Elements of Classroom Management: Managing Time and Space, Student Behavior, and Instructional Strategies. ASCD.

Merriënboer, J. J. G. van, & Sweller, J. (2005). Teori Beban Kognitif dan Pembelajaran Kompleks. Tinjauan Psikologi Pendidikan.

Miftah, M. (2022). Efektivitas Pemanfaatan Media Berbasis TIK untuk Optimalisasi Pembelajaran. Publica Indonesia Utama.

Moreno, R. A., & Mayer, R. E. (2004). Personalized Messages that Promote Science Learning in Virtual Environments. Journal of Educational Psychology, 96(1), 165–173.

https://doi.org/https://doi.org/10.1037/0022-0663.96.1.165

Nafadilla, Panjaitan, K. R., Ayu, M. S., & Claudia, A. D. (2025). Efisiensi Penggunaan Waktu Pembelajaran sebagai Implementasi Manajemen Pendidikan di Kelas Sekolah Dasar Efficient Use of Learning Time as an Implementation of Educational Management in Elementary School Classes. JIIC: Jurnal Intelek Insan Cendikia, 11687–11692.

Noetel, M., Griffith, S., Delaney, O., Harris, N. R., Sanders, T., Parker, P., & del Pozo Cruz, B., Lonsdale, C. (2022). Multimedia Design for Learning: An Overview of Reviews with Meta-meta-analysis. Review of Educational Research, 92(3), 413–454. https://doi.org/https://doi.org/10.3102/00346543211052329

Ouwehand, K., Kroef, A. van der, Wong, J., & Paas, F. (2021). Measuring Cognitive Load: Are There More Valid Alternatives to Likert Rating Scales? Frontiers in Education, 6, 702616.

Paas, F. (1992). Training Strategies for Attaining Transfer of Problem-Solving Skill in Statistics: A Cognitive Load Approach. Journal of Educational Psychology, 84(4), 429–434. https://doi.org/https://doi.org/10.1037/0022-0663.84.4.429

Paas, F. (2010). Konseptualisasi Baru dalam CLT. Tinjauan Psikologi Pendidikan.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive Load Theory and Instructional Design: Recent Developments. Educational Psychologist.

Prince, M. J., Vigeant, M., & Nottis, K. E. K. (2012). Pengembangan Inventaris Konsep Panas dan Energi. Jurnal Pendidikan Teknik.

Schuessler, K., Fischer, V., & Walpuski, M. (2025). Investigating Construct Validity of Cognitive Load Rating Scales. Instructional Science. https://doi.org/https://doi.org/10.1007/s11251-024-09692-6

Schulte-Mecklenbeck, M., Kühberger, A., & Johnson, J. G. (2019). A Handbook of Process Tracing Methods. Routledge, New York, NY.

Sweller, J. (1988). Cognitive Load During Problem Solving: Effects on Learning. Cognitive Science, 12, 257–285.

Trypke, M., Stebner, F., & Wirth, J. (2023). Two Types of Redundancy in Multimedia Learning: A Literature Review. Frontiers in Psychology. https://doi.org/https://doi.org/10.3389/fpsyg.2023.1148035

Tseng, C. H., Chao, C.-J., & Lin, S.-F. (2023). Panas & Suhu Melalui Laboratorium Fisik vs. Virtual. Journal Pendidikan Sains Baltik.

UNICEF Indonesia. (2021). Strengthening Digital Learning across Indonesia : A Study Brief. Unicef. https://www.unicef.org/indonesia/media/10531/file/StrengtheningDigitalLearningacrossIndonesia%3AAStudyBrief.pdf

UNICEF Indonesia. (2023). Analisis Situasi untuk Lanskap Pembelajaran Digital di Indonesia. Unicef. https://www.unicef.org/indonesia/media/13421/file/AnalisisSituasiuntukLanskapPembelajaranDigitaldiIndonesia.pdf

Wang, C., Zhang, M., Sesunan, A., & Yolanda, L. (2023). Peran Teknologi dalam Transformasi Pendidikan di Indonesia. Kemdikbud, 4(2), 1–7.

Zacharia, Olympiou, G., & Papaevripidou, M. (2008). Argumen untuk Menggabungkan Laboratorium Fisik dan Virtual. Jurnal Penelitian Penngajaran Sains.

Zhu-Tian, C., & Xia, H. (2022). Cross Data: Leveraging Text-Data Connections for Authoring Data Documents. Proceedings of the 2022 CHI Conference on Human Fact*ors in Computing Systems*, 1–15.