



AR Interactivity for Environmental Learning: A Cross-Case Study of Eco-Visualization

Muhammad Nur Hudha*

Universitas Sebelas Maret,
INDONESIA

Riezky Maya Probosari

Universitas Sebelas Maret,
INDONESIA

Kadek Dwi Hendratma Gunawan

Universitas Sebelas Maret,
INDONESIA

Nada Dhenia Febriyanti

Universitas Sebelas Maret,
INDONESIA

Dwi Nursanti

Universitas Sebelas Maret,
INDONESIA

Firmanul Catur Wibowo

Universitas Negeri Jakarta,
INDONESIA

Article Info

Article history:

Received: June 11, 2026

Revised: September 9, 2026

Accepted: September 18, 2026

Keywords:

Augmented reality;

Eco-visualization;

Education for sustainable
development;

ICAP framework;

Interactivity.

Abstract

Environmental concepts such as pollutant dispersal and renewable-energy systems are difficult to observe directly, creating challenges for conceptual understanding and ecological awareness in secondary science. Although augmented reality (AR) can visualize these phenomena, limited guidance exists on how its interactivity should be aligned with different learning goals. Drawing on the ICAP framework and cognitive theory of multimedia learning, this study examined interactivity–outcome alignment through a multiple-case Research and Development design using ADDIE. Two contrasting media were developed and evaluated: low-interactivity AR flashcards for pollution targeting ecological awareness and a higher-interactivity PhET–AR flipped simulation for renewable energy targeting conceptual understanding. Cross-case analysis showed that both media were considered valid and practical and produced meaningful improvements in their targeted learning outcomes. Ecological-awareness gains were more evident in knowledge than in pro-environmental behavioral intention, while the simulation-supported case improved conceptual understanding. Rather than establishing a causal advantage of one interactivity level over another, the findings suggest that AR effectiveness depends on aligning learner interaction with the intended outcome and appropriate instructional scaffolding. The cross-case synthesis generates six testable design propositions for developing goal-aligned AR eco-visualization and provides a foundation for further comparative research in technology-enhanced science learning.

To cite this article: Hudha, M. N., Probosari, R. M., Gunawan, K. D. H., Febriyanti, N. D., Nursanti, D., & Wibowo, F. C. (2026). AR interactivity for environmental learning: A cross-case study of eco-visualization. *Online Learning in Educational Research*, 6(3), 201-217. <https://doi.org/10.58524/oler.v6i3.1436>

INTRODUCTION

Education for Sustainable Development (ESD) frames schools as an important site for cultivating environmental literacy and sustainability competencies, and curriculum policy in many countries, including Indonesia, increasingly assigns them that role (Annan-Diab & Molinari, 2017; Kioupi & Voulvoulis, 2019; Taufik et al., 2026). Yet such content remains thinly represented in school science materials (Hudha et al., 2021), and in secondary science many environmentally relevant concepts are intrinsically difficult to learn because they are abstract, span spatial and temporal scales students cannot perceive, and unfold as dynamic processes rather than static facts. Pollutant movement through air, water, and soil, and energy conversion in renewable-energy systems, are typical examples.

* Corresponding Author

Muhammad Nur Hudha, Universitas Sebelas Maret, INDONESIA ✉ hudha.mn@staff.uns.ac.id

Augmented reality (AR) addresses precisely this representational problem. By overlaying contextual, three-dimensional, and often interactive virtual content onto the physical world, AR renders abstract structures and processes as spatial, frequently animated representations that learners can inspect from multiple viewpoints (Akçayır & Akçayır, 2017; Santos et al., 2014; Wu et al., 2013). Reviews of AR in science education report consistent benefits for conceptual understanding, motivation, and engagement (AlGerafi et al., 2023; Anesha et al., 2026; Saputra et al., 2026). Meta-analyses place the average benefit for achievement in the medium range (Garzón et al., 2019; Kalemkuş & Kalemkuş, 2022; Küçük Avcı et al., 2019; Li et al., 2021). In environmental education specifically, AR interventions have improved learners' knowledge of, attitudes toward, and awareness of environmental issues across school levels, suggesting that the technology can serve both the cognitive and the affective aims of ESD (Attanasi et al., 2025; Castillo-Gonzalez et al., 2023; Khan et al., 2026; Nugroho et al., 2025; Ozturk & Akcay, 2023).

These media are not pedagogically uniform, however, and they vary in how much learner agency they afford. At the lower-interactivity end, marker-triggered AR primarily supports recognition and recall: learners scan a target and view an explanatory 3D object or animation (Tzima et al., 2019). At the higher-interactivity end, simulation-based AR supports inquiry, because learners manipulate variables and observe consequences, as in PhET interactive simulations, which repeatedly strengthen conceptual understanding by linking macroscopic, symbolic, and sub-microscopic representations (Furqon, 2024; Pranata, 2023; Rahmawati et al., 2022), by enabling scientific-discovery learning (de Jong & van Joolingen, 1998), and through inquiry-oriented designs that support conceptual understanding and science literacy (Aviyanti et al., 2025; Hudha et al., 2019). We treat this variation as an interactivity continuum running from recognition-oriented to manipulation-oriented designs. The goals these media serve also vary. ESD requires both cognitive outcomes, meaning accurate conceptual understanding of environmental systems, and affective outcomes, typically theorized as ecological awareness encompassing knowledge, sensitivity, and pro-environmental behavioral intention. These outcome types are not interchangeable. A well-documented knowledge-attitude-behavior gap means that instruction often raises environmental knowledge more readily than it shifts behavioral intention (Debrah et al., 2021; Kollmuss & Agyeman, 2002).

Despite the volume of this work, three limitations constrain its usefulness to designers. First, evaluations are typically single-product, leaving unclear which design features transfer beyond the artifact evaluated (Anesha et al., 2026; Saputra et al., 2026). Second, studies generally target a single outcome, while synthesis literature treats achievement and environmental outcomes separately, leaving unclear whether one design can serve both (Castillo-Gonzalez et al., 2023; Garzón et al., 2019; Kalemkuş & Kalemkuş, 2022; Küçük Avcı et al., 2019; Li et al., 2021; Ozturk & Akcay, 2023). Third, development studies commonly establish that a product is valid, practical, and effective without theorizing why particular design choices produced those results (Saputra et al., 2026; Wang & Li, 2024). Reviews attribute heterogeneity in AR effects to moderators such as duration, discipline, and device (Garzón et al., 2019; Li et al., 2021), but the fit between interactivity and outcome type has received less explicit attention. This study therefore proposes a provisional framework linking these two dimensions, intended to generate hypotheses rather than establish an empirically confirmed relationship.

The interactive-constructive-active-passive (ICAP) framework (Chi & Wylie, 2014) orders four modes of overt learner engagement, from passive through active and constructive to interactive, and predicts that learning outcomes improve across that order. The two design families map onto it as follows. Recognition-oriented AR primarily involves active engagement with prepared representations. In contrast, manipulation-oriented AR can support constructive engagement by requiring learners to set variables, predict outcomes, observe consequences, and reconcile discrepancies. Because ICAP was formulated primarily to explain knowledge and comprehension, it does not by itself establish effects on attitudes. Accordingly, this study treats the knowledge component of ecological awareness as the relevant affective-related outcome that ICAP can plausibly inform, while treating sensitivity and behavioral intention as outcomes to be examined rather than predicted by the framework.

The cognitive theory of multimedia learning (Mayer, 2014) further explains why spatial and dual-channel representations may support learning by reducing unnecessary processing and

supporting the construction of coherent mental models. Cognitive load theory (Sweller et al., 2019) adds that highly manipulable simulations may impose greater processing demands on novice learners because they must coordinate multiple elements and interactions simultaneously. Accordingly, high-interactivity designs may require instructional scaffolding, such as worksheets or flipped pre-work, whereas recognition-oriented AR may impose lighter processing demands.

Taken together, these accounts motivate a prediction: low-interactivity AR should efficiently match affective and knowledge goals, whereas high-interactivity AR, suitably scaffolded, should better match cognitive-restructuring goals. We state this as a prediction to be examined. It is not an established consequence of the theories, since neither ICAP nor cognitive load theory was developed to adjudicate between media at different points on an interactivity continuum. Crossing that continuum with outcome type yields the design space shown in Figure 1.

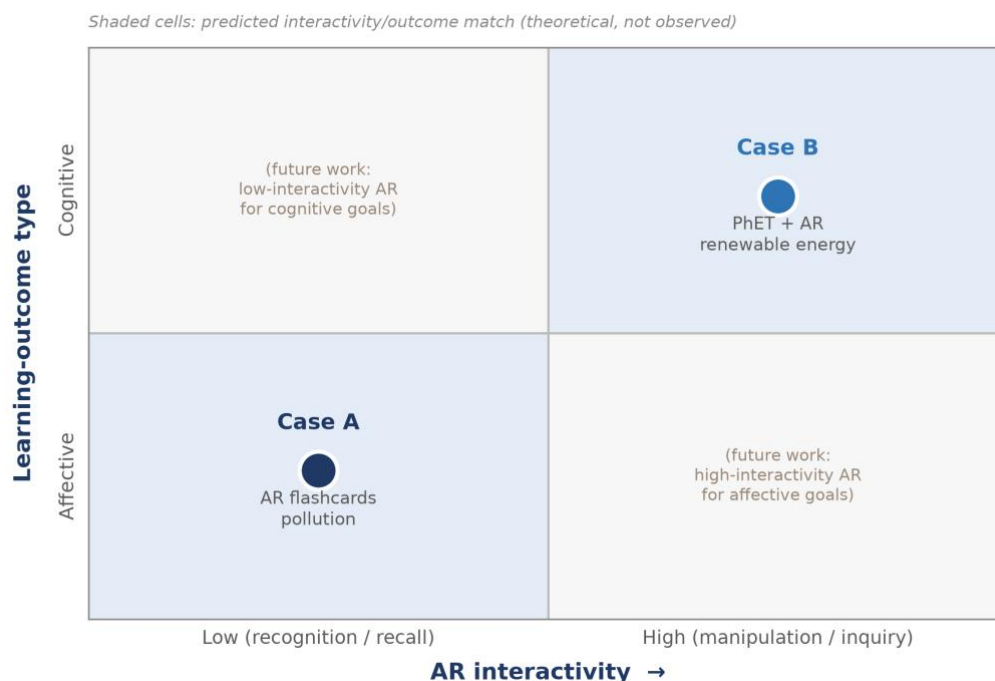


Figure 1. Space Crossing AR Interactivity Design with Learning-Outcome Type

The diagonal represents a theoretical proposition rather than an empirical result; the present cases illustrate this proposed alignment but do not test it independently. This study populates that space by analyzing two AR developments as complementary cases at different points on the interactivity continuum. The contribution is threefold. Theoretically, we propose a framework relating AR interactivity to outcome type, grounded in ICAP and multimedia-learning accounts, not in product evaluation. Methodologically, we introduce a cross-case re-analysis procedure that decomposes an affective outcome into its dimensions and examines the distribution, not only the mean, of a cognitive gain, yielding evidence on the internal structure of learning gains that single-outcome reports cannot provide. Practically, we derive six design propositions for AR eco-visualization that are feasible in low-infrastructure school settings. Three research questions guide the study:

- RQ1: How feasible (valid) and practical are AR eco-visualization media developed with ADDIE for secondary-science environmental topics?
- RQ2: How effective are low-interactivity and high-interactivity AR media in improving ecological awareness and conceptual understanding, respectively, and what is the internal structure of those gains?
- RQ3: What transferable design propositions emerge from a cross-case comparison, and how do they relate AR interactivity to outcome type?

METHOD

Design

The primary design is a multiple-case Research and Development study. Each case is a complete ADDIE cycle (Analysis, Design, Development, Implementation, Evaluation) for one AR medium (Branch, 2009). ADDIE governed the original development and evaluation of each medium, while the present article adds a cross-case analysis of the two completed cycles. We do not claim design-based research in the formal methodological sense because each case involved a single development cycle rather than iterative intervention and revision with practitioners. The contribution is therefore framed as analytic rather than statistical generalization to design theory (Yin, 2018).

The cross-case procedure followed four steps, adapted from established protocols for multiple-case designs (Handhika et al., 2025; Miles et al., 2014; Yin, 2018). Each case was first written up independently on a fixed template covering context, design features, validation evidence, practicality evidence, and effectiveness evidence. The two accounts were then entered into a common display whose rows are the evaluation dimensions shared by both reports. Effectiveness was recomputed for both cases on a single procedure, so that the comparison rests on identically derived statistics. Finally, features that recurred across the cases, or differed systematically between them, were abstracted into candidate propositions, each checked back against the within-case accounts and retained only where both cases bear on it. We selected the cases as an information-rich pairing because they share method, technology platform, and analytic instruments while differing in interactivity and outcome type.

Because that continuum carries the argument of this article, we state the criteria by which a medium is placed on it rather than relying on the labels alone. Table 1 sets out four dimensions and the engagement mode each placement implies. The dimensions adapt distinctions from learner-agency and simulation-based-learning accounts into criteria usable for coding an AR medium (Chi & Wylie, 2014; de Jong & van Joolingen, 1998; Suyidno et al., 2025). Placement on the continuum is a matter of degree, so a medium may sit near an end without occupying it.

Table 1. Criteria for Placing an AR Medium on the Interactivity Continuum, with the ICAP Mode Each Placement Implies

Dimension	Recognition-Oriented (low)	Manipulation-Oriented (high)
Action on content	Trigger and inspect a prepared representation	Set parameters that alter the represented system
Alteration of system	Viewpoint changes; represented content invariant	System state changes with learner input
Response contingency	Same output on every trigger	Output varies with the input chosen
Learner control	Sequence and pacing	Sequence, pacing, and variable values
<i>Implied ICAP mode</i>	<i>Active</i>	<i>Constructive, becoming interactive with partner work</i>

Note. Rows 1–4 are the classification criteria. The last row, in italics, is not a criterion but the ICAP mode each placement implies.

The two media were then coded against the criteria in Table 1. Case A triggers a prepared 3D scene from a marker: the learner changes viewpoint but not the represented content, every trigger returns the same scene, and control extends only to sequence and pacing. Case B sets variables in a simulated energy system: system state changes continuously with input, output is contingent on the settings chosen, and control extends to the variable values themselves. Case A is not purely recognition-oriented, since its worksheet requires a choice between two courses of action and then displays the consequence; because that choice is a single scripted branch and not parametric manipulation, the medium sits at the low end of the continuum rather than at its endpoint.

Cases, Participants, and Ethics

Both cases ran in the even semester of the 2025/2026 academic year and used the same two-stage trial structure. A small-scale usability trial with 10 students came first, followed by a limited

field trial with one intact class. The Case B study has since been published separately (Febriyanti et al., 2026); this article re-analyses its raw pretest and posttest scores using a procedure common to both cases.

Case A was conducted at a state junior secondary school in Sragen Regency, Central Java, that holds Adiwiyata (environmentally focused school) status. The limited field trial used one intact Grade VII class ($n = 30$; age approximately 12–13 years). Case B was conducted at a state junior secondary school in Karanganyar Regency, Central Java. The limited field trial used one intact Grade IX class ($n = 34$; age approximately 14–15 years).

In both cases, the class was selected by convenience sampling from classes at the target grade based on practical criteria reported in the development studies: prerequisite content had been covered, reliable internet access was available, smartphone use was permitted, and sufficient devices were available for group work. Case B additionally required willingness to participate, no visual impairment that would prevent use of the visual media, and completion of the response questionnaire. Sampling was therefore non-probability and purposive, and no class was assigned to a condition; each case was evaluated in the setting for which its medium was designed.

We did not select the two grade levels for comparison. Still, they followed curriculum placement, with environmental pollution taught in Grade VII and renewable energy in Grade IX under the Indonesian Merdeka Curriculum. Thus, grade level, school, and topic are confounded with case condition, limiting attribution of cross-case differences to interactivity alone. We obtained school permission, teacher and parent/guardian consent, and student assent. Participation was voluntary, and we anonymized student records before analysis.

The Two AR Media

Case A (Low Interactivity)

AR-enhanced flashcards covering four pollution topics (general environmental pollution, air, soil, water). Cards were designed in Canva and printed on heavyweight matte stock to support reliable marker detection; AR scenes were authored in Assemblr Edu with 3D objects generated using a text-to-3D tool, and were paired with a worksheet so that scanning a marker led students through a short scenario, a decision, and a reflection on everyday pro-environmental behavior.

Case B (High Interactivity)

A package combining PhET virtual-laboratory simulations with AR 3D visualizations authored in Assemblr Edu, delivered through a flipped-classroom approach in which students engaged with pre-class video and used class time for manipulation and discussion. A second, context-specific AR artifact (“Environmental Cases”) anchored renewable-energy concepts in real situations by contrasting environments degraded by fossil-fuel emissions with environments served by renewable energy. Figure 2 locates the two media along the interactivity continuum.

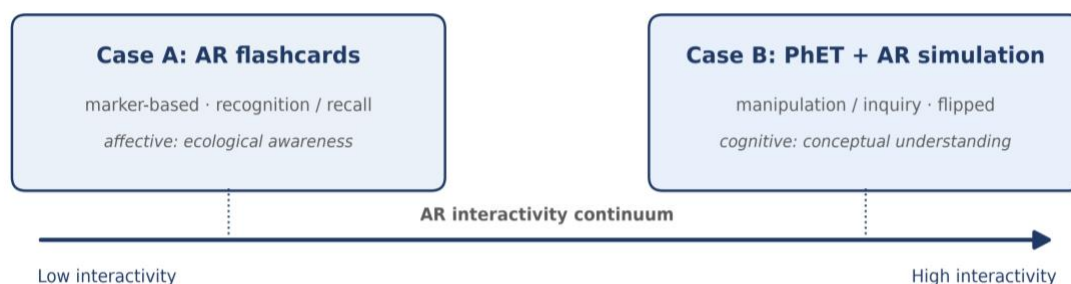


Figure 2. Schematic of the Two AR Media along the Interactivity Continuum

Implementation Procedure

In Case A, students worked in small groups over four 80-minute lessons (2×40), one lesson per pollution topic. Each cycle began with a briefing on marker scanning, after which students scanned a flashcard marker with their own smartphone to open the corresponding Assemblr Edu scene. The scene posed a situated question requiring students to choose a course of action; groups

then discussed the consequences of the chosen action for the environment and for living things, consulted a short explanatory text embedded in the AR scene, and closed with a reflection on their own everyday behavior. A validated worksheet structured this sequence across the four pollution topics, and a printed user guide supported independent operation of the media. We administered the posttest after the final cycle.

In Case B, the intervention ran across five 80-minute meetings (2×40) in a flipped-classroom format. Before each meeting, students worked independently at home with an introductory video on renewable energy and a tutorial on operating the PhET and AR media, accessed through QR codes printed on the worksheet. Class time followed a three-stage sequence. A short warm-up activity first activated what students had studied at home. The teacher then clarified concepts the video left unclear and established a common baseline. In the final stage, students manipulated the PhET simulations and AR scenes individually and in groups, while the teacher circulated as a facilitator and provided immediate feedback. The pretest and posttest were administered before the first and after the final meeting.

Instruments and Their Validation

Expert Raters and Rating Procedure and Interpretation

Two expert raters judged each medium and rated every validation domain. Both hold doctorates in science education and have expertise in technology-mediated science teaching, augmented reality and physics instruction, multimedia and simulation-based learning, and education for sustainable development. Both raters co-authored the present article; therefore, they produced the ratings within the development teams rather than through an external panel. The Gregory index is consequently treated as evidence of agreement on item relevance, not as independent external appraisal or construct validity.

Each rater judged every item on the validation sheet as relevant or not relevant. Agreement was quantified with the Gregory content-validity index, $D/(A+B+C+D)$, where D counts items both raters judged relevant, B and C the items only one judged relevant, and A the items neither did. Indices were read against the conventional bands used in both development reports: 0.80–1.00 very high, 0.60–0.79 high, 0.40–0.59 moderate, 0.20–0.39 low, below 0.20 very low. We retained both media for further analysis only when all validation domains reached the high or very high category.

Case A Instruments

The media validation sheet carried 10 items across five aspects (fit to the lesson, screen design, text, image quality, software operation); the content sheet carried 14 items across curriculum fit, accuracy, completeness, order of presentation, and language. Ecological awareness was measured with a 30-item multiple-choice test, scored dichotomously and written against the three dimensions of the construct. Its content validity was 0.93 (pretest form) and 0.90 (posttest form). Item construct validity was then examined against the product–moment critical value; 25 of 30 pretest items and 26 of 30 posttest items were retained, so the forms differ slightly in length, and each score is a percentage of the items retained on the form taken. Internal consistency of the retained items, by the Kuder–Richardson formula 20 (KR-20), was 0.876 and 0.869; item difficulty and discrimination were computed for both forms. An 18-item affective observation sheet, completed by two observers during the lessons and across the following week, triangulated the test. Practicality used four-point Likert questionnaires, 9 indicators for the teacher and 12 for students, each spanning ease of use, usefulness, and satisfaction.

The Case A test was blueprinted against the three dimensions of ecological awareness and their sub-indicators: ecological knowledge (knowing how human behavior affects the environment; understanding why cleanliness and ecological balance matter), ecological sensitivity (concern for environmental conditions, empathy toward living things, emotional attachment to nature, and sensitivity to local environmental problems), and pro-environmental behavior (minimizing resource use; reducing waste production). Items were distributed across these sub-indicators, and the dimension-level analysis aggregates the retained items for each: 8, 10, and 7 items on the pretest form, and 8, 9, and 9 on the posttest form, which reconciles the two forms to 25 and 26 retained items.

Case B Instruments

The media validation sheet carried 18 statements on graphic quality plus one supporting question; the content-and-language sheet carried 32 statements across content feasibility (10), presentation (6), language (11), and contextual appropriateness (5), plus four supporting questions. We measured conceptual understanding with a 25-item multiple-choice test, scored dichotomously and mapped to seven understanding indicators. Item construct validity was examined using the product-moment critical value ($r = 0.339$ at $n = 34$); all 25 items exceeded the threshold on both the pretest and posttest forms, so none were discarded, and each student's score is the percentage of the 25 items answered correctly. Practicality was measured with four-point Likert response questionnaires for students and the subject teacher.

The Case B test was blueprinted against seven indicators of conceptual understanding, corresponding to the cognitive processes of the understand category: interpreting, exemplifying, classifying, summarizing, inferring, comparing, and explaining. Two items illustrate these indicators. Targeting explaining: "The main role of the generator in almost all renewable-electricity plants (hydro, wind, geothermal) is to ...", keyed to "convert the mechanical energy of the rotating turbine into electrical energy". Targeting inference: "Rooftop solar is popular in cities, whereas hydroelectric plants are built in rural areas near rivers. From these facts it can be concluded that ...", keyed to "the choice of renewable-energy type is strongly influenced by local natural-resource potential".

Data Analysis

In the original development studies, practicality was reported as percentage scores, and effectiveness was evaluated using Hake's normalized gain, $\backslash(g = (\text{post-pre})/(100\text{-pre})\backslash)$, computed in IBM SPSS Statistics 27. For this cross-case analysis, we recomputed effectiveness in Python 3.12.7 using SciPy 1.17.0 and NumPy 2.5.1. The recomputed individual N-gains reproduced the original SPSS values to four decimal places (Case A: $M = 0.5227$, $SD = 0.2819$; Case B: $M = 0.4972$, $SD = 0.57441$).

For each case, we summarized pretest and posttest scores using means and standard deviations. We calculated paired-samples t-tests, Cohen's $\backslash(d_z\backslash)$, aggregate N-gain, per-student N-gain, and Pitman–Morgan tests. We assessed normality of paired differences using Shapiro–Wilk tests. Case A showed approximately normal differences ($W = 0.96$, $p = .31$), so we used the paired t-test as the primary inferential test and reported the Wilcoxon signed-rank test as a robustness check. Case B violated normality ($W = 0.73$, $p < .001$; skewness = 2.70; excess kurtosis = 10.13), so the Wilcoxon signed-rank test was primary, with matched-pairs rank-biserial correlation as the distribution-free effect size. We retained the paired t-test and $\backslash(d_z\backslash)$ for comparability with prior AR and simulation studies. We calculated 95% confidence intervals for mean per-student N-gain using t-based intervals. For Case A, we also decomposed N-gain across the three ecological-awareness dimensions using item-level records for all 30 students.

Two cautions guided interpretation of N-gain. First, Hake's N-gain was developed for conceptual-achievement data; its use for Case A's knowledge-referenced ecological-awareness measure is therefore descriptive, while inferential conclusions rely on the paired tests and effect sizes. Second, high pretest scores can make N-gain unstable because of limited headroom. This issue was relevant to Case B; therefore, we reported aggregate and per-student N-gain together. Because the bounded percentage scale may also affect variance comparisons, we additionally examined pretest-to-posttest variance change on logit and arcsine-transformed scales.

RESULTS AND DISCUSSION

Results are organized by research question. Sections 1 and 2 report the two cases in turn: within each, the validation and practicality evidence answer RQ1 and the pretest–posttest analysis answers RQ2, with the statistical results presented first and their interpretation following in a separate paragraph. Section 3 answers RQ3 by placing the cases in a common frame and deriving the design propositions.

Read together, the pattern is compatible with the framework's central prediction, namely that the appropriate level of AR interactivity depends on the type of learning goal: a low-interactivity, recognition-oriented design was followed by movement on an affective outcome, and a cognitive

outcome was addressed with a high-interactivity, manipulation-oriented design. Because the two cases differ in grade, school, topic, duration, and supporting pedagogy as well as in interactivity and outcome, and because neither case included a control group, this compatibility is a design inference drawn across two development cycles rather than an experimentally isolated effect. We therefore state it throughout as a proposition to be tested, not as a demonstrated cause.

Case A: AR Flashcards for Ecological Awareness

Feasibility and Practicality (RQ1)

Expert validation indicated high agreement on both aspects (media-aspect Gregory index = 1.00 across the 10 media items; content-aspect index = 0.86 across the 14 content items), so the medium was feasible for trial. Practicality was high from both perspectives, with the teacher response at 80% and the student response at 82.29%, both in the “good/high” band.

Effectiveness and the Structure of the Gain (RQ2)

Re-analysis of the raw scores ($n = 30$) substantiates a reliable affective improvement: the mean rose from 67.20 (SD = 13.35) to 85.07 (SD = 8.51), with a median gain of 20.0 points. Because the paired differences were approximately normal (Shapiro–Wilk $W = 0.96$, $p = .31$), the increase was tested with a paired-samples t -test and is highly reliable: $t(29) = 9.02$, $p < .001$, $d_z = 1.65$, with the Wilcoxon signed-rank test agreeing ($p < .001$). The per-student N-gain reproduced the original value ($M = 0.52$, SD = 0.28, 95% CI [0.42, 0.63]; per-student values ranged from -0.33 to 1.00), and the classical aggregate N-gain (0.54) was nearly identical. Score variance also fell from pretest to posttest (SD 13.35 $\rightarrow$ 8.51), a significant compression (Pitman–Morgan $t(28) = 3.04$, $p = .005$), and an affective observation score of 73% corroborated the test-based gain.

The original study reported gains across all three dimensions of ecological awareness; our decomposition into dimension-level normalized gains (Table 2, Figure 3a) reveals a clear gradient: the medium moved ecological knowledge most ($\langle g \rangle = 0.69$), sensitivity next (0.60), and pro-environmental behavior least (0.46). These differential gains across dimensions align with what the knowledge–attitude–behavior gap would suggest, and they are a pattern the overall score conceals. The three dimensions are components of ecological awareness measured at the same two time points, not sequential stages tracked longitudinally. Hence, the gradient describes where the medium had most and least purchase, not a developmental progression from knowing to acting. The overall N-gain (per-student $M = 0.52$) is computed on total awareness scores and therefore need not equal the mean of the three dimension-level gains (0.69, 0.60, 0.46), which are derived from dimension percentages.

Table 2. Case A: Dimension-Level Decomposition of the Affective Gain (New Analysis)

Dimension	Pretest	Posttest	Normalized gain $\langle g \rangle$
Ecological knowledge	70.0%	90.8%	0.69
Ecological sensitivity	69.7%	87.8%	0.60
Pro-environmental behavior	60.5%	78.5%	0.46

Case A's most informative result is not its medium aggregate gain but this internal gradient. It is a quantified, single-intervention instance of the well-documented knowledge–attitude–behavior gap (Fujii, 2006; Kollmuss & Agyeman, 2002), and it converges with environmental-education studies, including from Indonesian secondary schools, that repeatedly find knowledge rising more readily than behavioral commitment (Handayani & Widodo, 2024; Harding et al., 2018; Laaloua, 2026). It also adds nuance to AR-environmental-awareness studies that report aggregate awareness gains without decomposition (Attanasi et al., 2025; Ozturk & Akcay, 2023), since the benefit of recognition AR appears concentrated at the knowledge end of the affective spectrum. Theoretically, the pattern sits at the boundary of what ICAP can speak to. Active, recognition-level engagement was enough to move the knowledge component, which is the kind of outcome ICAP was built to predict; sensitivity and behavioral intention lie outside its scope, and the gradient neither confirms nor disconfirms it there. Read that way, the pattern motivates Proposition 4 — that affective AR be paired

with explicit reflection-and-action tasks — as a design response to where the medium fell short, not as a prediction the framework supplied.

Case B: PhET-and-AR Simulation for Conceptual Understanding

Feasibility and Practicality (RQ1)

Expert validation was very high, with a media-aspect Gregory index of 0.89 and a content/language index of 1.00, and all conceptual-understanding items passed construct validity. Practicality was strong, with student responses of 91% (small-scale) and 81% (limited field trial) and a teacher response of 93%.

Effectiveness and the Structure of the Gain (RQ2)

Re-analysis of the raw scores ($n = 34$) substantiates a reliable cognitive improvement (Table 3): the mean rose from 81.18 ($SD = 16.38$) to 91.53 ($SD = 8.51$), with a median gain of 8.0 points (interquartile range [IQR] 4.0 to 15.0). Because the paired differences departed markedly from normality (Shapiro–Wilk $W = 0.73$, $p < .001$), we tested the increase with a Wilcoxon signed-rank test, which was highly reliable ($p < .001$). We showed a large distribution-free effect (rank-biserial $r = 0.84$). The parametric counterparts, reported for comparability with the meta-analytic literature, agree: $t(33) = 3.88$, $p < .001$, $dz = 0.66$.

The per-student N-gain reproduces the original value ($M = 0.497$, $SD = 0.574$, 95% CI [0.30, 0.69]); a classical aggregate N-gain of 0.55 is less sensitive to ceiling effects. Two features that the original report did not surface are notable. The gain is heterogeneous (35% high, 41% medium, 24% low or negative; per-student values range from -2.00 to 1.00 , and two students declined), reflecting ceiling effects among already-high pretest scorers. The score variance also fell from pretest to posttest ($SD\ 16.38 \rightarrow 8.51$), a statistically significant compression (Pitman–Morgan $t(32) = 4.24$, $p < .001$; Figure 3b). Descriptively, the posttest distribution is not a translation of the pretest distribution: initially low performers moved further than initially high ones, whose scores were already close to the maximum.

Table 3. Case B: Re-analysis of Pretest–Posttest Data ($n = 34$; New Analysis)

Statistic	Value
Pretest mean (SD)	81.18 (16.38)
Posttest mean (SD)	91.53 (8.51)
Median gain (IQR)	8.0 (4.0–15.0)
Normality of paired differences (Shapiro–Wilk)	$W = 0.73$, $p < .001$ (not met)
Wilcoxon signed-rank test	$p < .001$
Rank-biserial correlation	0.84 (large)
Paired t-test	$t(33) = 3.88$, $p < .001$
Within-subject effect size (Cohen's d)	0.66 (medium–large)
N-gain, per-student: M (SD)	0.497 (0.574)
N-gain, per-student: 95% CI	[0.30, 0.69]
N-gain, per-student: range	-2.00 to 1.00
N-gain, classical aggregate	0.55 (medium)
Gain distribution (high/medium/low-neg)	35% / 41% / 24%
Variance change (Pitman–Morgan)	$t(32) = 4.24$, $p < .001$ (compression)

Reporting the distribution, and not only the mean, changes what the result can be said to show. The spread narrowed in both cases, so it is not tied to manipulation-oriented design and, at least partly, to the measurement scale: with posttest means of 85.07 and 91.53 on a 0–100 scale, the upper bound constrains how much spread can remain. Re-tested on the logit and arcsine scales, which lift that bound, the compression is no longer detectable in either case (Case A $p = .38$ and $.50$; Case B $p = .46$ and $.66$), though with $n = 30$ and 34 these tests cannot establish its absence. We therefore report the narrowing as a property of the observed distributions, not as evidence that AR relieved cognitive load or delivered an equity benefit. Load was never measured, so accounts predicting such an effect (Keller et al., 2021; Suo et al., 2026; Zhu et al., 2026) remain untested here. The analysis does

establish one methodological point: distributional checks belong alongside mean effect sizes, since an apparent design effect need not survive a change of scale.

The magnitude of the gain is best read against simulation and AR benchmarks rather than as a stand-alone figure. The aggregate N-gain of 0.55 sits just above Hake's (1998) classic threshold separating traditional courses ($\langle g \rangle \approx 0.23$) from interactive-engagement courses ($\langle g \rangle \approx 0.48$), which is compatible with a design that is, by construction, strongly interactive: a manipulable simulation embedded in a flipped, discussion-rich format. It is also consonant with meta-analytic evidence that computer simulations improve K-12 conceptual understanding (Hu et al., 2026) and that PhET in particular yields large gains in physics understanding (Khofifah et al., 2024; Suherman et al., 2025), with flipped-classroom meta-analyses reporting small-to-moderate advantages over conventional instruction (Van alten et al., 2019; Yu & Yu, 2024), and with blended, video-supported designs in comparable Indonesian settings (Huda et al., 2018).

Even so, interpret the within-subject effect size ($d_z = 0.66$) cautiously. Within-subject, one-group effect sizes are generally larger than the between-group estimates that dominate AR meta-analyses, where pooled effects typically fall in the medium range by Cohen's (1988) conventions (Garzón et al., 2019; Kalemkuş & Kalemkuş, 2022; Küçük Avcı et al., 2019; Li et al., 2021, $d \approx 0.46$) and only occasionally reach the large values reported for recent secondary-science AR (Barcenas & Prudente, 2026). Our medium magnitude, obtained without a control group, is therefore plausibly an upper bound and should not be read as evidence of a large, controlled effect. The gains are best read as real and consistent with the field, not exceptional. That is why the contribution rests on the structure of the gains, not their size.

Cross-Case Synthesis and Design Propositions

This section answers RQ3 by returning to RQ1 and RQ2 across both cases. Both media met their feasibility and practicality criteria. They produced medium-sized gains, but the gains are not structured alike: only the affective case shows a marked spread across its outcome components. RQ3 therefore asks what, if anything, transfers from two cycles that agree on magnitude and differ in internal structure.

Table 4 places the cases on the common evaluation frame. The pattern is convergent at the aggregate level, with both media reaching high feasibility and practicality and producing medium effectiveness despite sharply different interactivity and target outcomes. The additional analyses reveal two recurring features. Both cases show a narrowing of score spread from pretest to posttest on the percentage scale (Pitman-Morgan $t(28) = 3.04$, $p = .005$ for Case A; $t(32) = 4.24$, $p < .001$ for Case B). The scale-transformation checks reported above attribute that narrowing largely to the scale's upper bound, not to the design. The affective case alone shows differential gains across its three dimensions, largest for knowledge and smallest for behavior (Figure 3). Only the second is specific to an outcome type, and it is the firmer of the two. Read together with the similar aggregate magnitudes, these features are compatible with the interactivity-outcome framework and motivate the design propositions that follow; they do not, in two cases that differ in several respects at once, establish it.

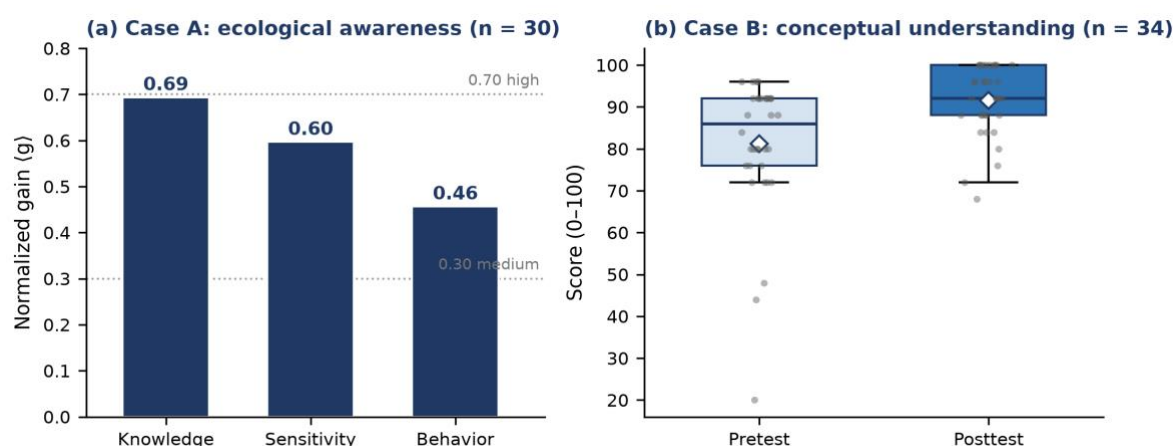


Figure 3. New Cross-Case Analyses: (a) Case A Affective Gain by Dimension; (b) Case B Score Distribution

Figure 3 presents the new cross-case analyses for the two cases, with $n = 30$ for Case A and $n = 34$ for Case B. In panel (a), Knowledge, Sensitivity, and Behavior refer to ecological knowledge, ecological sensitivity, and pro-environmental behavior, respectively. The reported values represent class-aggregate gains calculated from dimension percentages rather than means of individual-level gains, while the dotted lines indicate the conventional thresholds. Panel (b) presents the score distributions for Case B, where the boxes represent the interquartile ranges, horizontal lines indicate the medians, diamonds indicate the means, and grey points represent individual scores. The standard deviation (SD) for Case B decreased from 16.38 to 8.51, while the corresponding values for Case A were 13.35 and 8.51, respectively.

Panel (a) shows a monotone descent across the three dimensions. The steepest single step falls between sensitivity and behavior, from 0.60 to 0.46; the step from knowledge to sensitivity, 0.69 to 0.60, is shallower. The gradient is therefore not a uniform attenuation. The medium had the least purchase on the behavioral dimension, where the knowledge–attitude–behavior gap is conventionally located. As noted for Case A, the ordering describes how far the medium reached across components; it does not track students moving along a path from knowing to acting. Panel (b) makes the mechanism visible: the upper boundary is nearly fixed by the ceiling while the lower boundary moves substantially, so the bottom of the distribution drives the narrowing and the top contributes almost nothing. Together, the two panels show why only the first feature is specific to an outcome type. To summarize similarities and differences across feasibility, practicality, effectiveness, and the internal structure of the gains, Table 4 presents a cross-case comparison.

Table 4. Cross-Case Comparison on the Feasibility–Practicality–Effectiveness Frame

Dimension	Case A (Low Interactivity)	Case B (High Interactivity)
Target outcome	Ecological awareness (affective)	Conceptual understanding (cognitive)
Topic/grade	Environmental pollution / VII	Renewable energy / IX
AR role	Recognition/recall of 3D impacts	Manipulation/inquiry of systems
Feasibility (Gregory)	Media 1.00; content 0.86	Media 0.89; content/language 1.00
Practicality	Teacher 80%; student 82.29%	Student 91%/81%; teacher 93%
Effectiveness (N-gain)	0.52 per-student; 0.54 aggregate	0.497 per-student; 0.55 aggregate
Inferential test (new)	$t(29) = 9.02, p < .001, dz = 1.65$	Wilcoxon $p < .001$; $t(33) = 3.88, p < .001, dz = 0.66$
Internal structure (new)	Differential gains: knowledge > sensitivity > behavior; spread narrows on percentage scale	Spread narrows on percentage scale.

The two cases also speak to a persistent puzzle in the AR literature, namely the very high heterogeneity of reported effects (Barcenas & Prudente, 2026; Garzón et al., 2019; Li et al., 2021). To the moderators that reviews usually invoke, these cases add a candidate at the level of design: the alignment between an AR medium's interactivity and the type of learning goal, which the graded-engagement logic of ICAP (Chi & Wylie, 2014) and evidence from game-based science learning (Tsai & Tsai, 2020) make plausible. The framework also tempers technological optimism: neither medium produced a large effect, both showed substantial within-group heterogeneity, and the cognitive case was built on a structuring pedagogy (flipped pre-work) intended to manage the load that interactivity imposes (Sweller et al., 2019). AR is thus best understood not as a panacea but as a representational affordance whose payoff depends on matching interactivity to purpose and embedding it in sound instructional design (Hamilton et al., 2020; Maroukhas et al., 2023).

Set against prior work, the picture is one of ordinary magnitudes and unusual structure. The aggregate gains are unremarkable, as Section 2 has already set out, and that is informative in itself, because this article's claim does not rest on effect size. The decomposition is the less common move: meta-analytic work summarizes outcomes as mean effects, whereas splitting the affective outcome into components showed that a single figure can conceal a gradient. What differs from the literature is what was examined, not what was found. The gradient may also owe something to the instrument,

since a knowledge-referenced measure will register the knowledge component most sensitively and the behavioral component least. Separating instrument from medium would require an awareness measure whose three components are calibrated independently, which neither development report provides.

The three accounts fare unevenly. ICAP is weakly supported: the active-mode design was followed by movement on the knowledge-oriented outcome without a constructive-mode design being required. On the point that matters most for ESD, it is neither supported nor refuted, since ICAP makes no prediction about sensitivity or behavioral intention, and that is precisely where Case A's gains were smallest. This marks a boundary of the framework, not a failure, and it suggests that affective AR design should rest on accounts of attitude and behavior change, not on engagement mode alone. The cognitive theory of multimedia learning is consistent with both cases. Still, it discriminates between neither, since both media are spatial and dual-channel by construction and neither case measured mental-model quality. Cognitive load theory received no test at all, because load was never measured; the claim that flipped pre-work managed the simulation's intrinsic load remains a design rationale. Measuring load directly, as the AR load literature now does routinely (Suzuki et al., 2023; Zhu et al., 2026), would convert that rationale into evidence. Six design propositions for AR eco-visualization emerged across the cases, stated as testable propositions rather than validated laws:

1. P1 (spatial externalization). Three-dimensional, viewpoint-flexible representations can externalize environmental phenomena that cannot be directly observed, making otherwise unobservable phenomena more accessible through AR-based representations.
2. P2 (device-aware delivery). Marker-based, smartphone-deliverable AR can keep infrastructure demands low while remaining compatible with typical school conditions.
3. P3 (interactivity–outcome alignment). Recognition-oriented AR may be sufficient for supporting affective awareness. However, whether manipulation-oriented AR is required for conceptual restructuring remains unsettled in the present synthesis, because none of the cases withheld manipulation-oriented interactivity from a cognitive goal. Accordingly, this proposition requires the most direct testing through a controlled study.
4. P4 (close the knowledge–behavior gap). Because affective gains were smallest on the behavioral dimension, AR should be paired with reflection and action tasks that explicitly target behavioral intention.
5. P5 (scaffold the load of interactivity). High-interactivity AR should be embedded within a structuring pedagogy, such as worksheets or flipped pre-work, to manage the cognitive load that interactivity may impose on novice learners.
6. P6 (layered validation). Development risk can be managed by sequencing expert feasibility assessment using Gregory, user-practicality assessment, and effectiveness assessment using N-gain.

Methodologically, the cross-case approach converts two contrasting development cases into potentially transferable, theoretically grounded design knowledge, and the additional analyses strengthen the empirical contribution without collecting new data. Practically, the propositions suit low-infrastructure contexts, since both media ran on students' smartphones using freely available authoring tools, addressing a constraint repeatedly flagged in developing-country AR research (Alalwan et al., 2020; Wu et al., 2013). For ESD, combining an affective and a cognitive case within one framework speaks directly to the field's dual aims (Kioupi & Voulvoulis, 2019).

LIMITATIONS

Several limitations qualify these claims. Each case involved a single ADDIE cycle, a modest sample ($n = 30$ and $n = 34$), a short intervention (four and five meetings, respectively), and no control group; therefore, the pre-post changes are indicative rather than confirmatory and cannot be attributed to AR alone. In Case B, conceptual gain may also reflect the PhET simulation, flipped pre-work, teacher scaffolding and discussion, or differences between the Grade IX renewable-energy and Grade VII pollution contexts. Because these factors were not varied independently of interactivity, the interactivity–outcome alignment remains a design inference rather than an experimentally

isolated effect. Case B's high pretest mean (81.18) and posttest mean (91.53) also limited score headroom and may have contributed to instability in per-student N-gain. Measurement limitations include content validity based on two co-author raters rather than independent external appraisal, uneven reliability documentation across cases, nonparallel Case A test forms (25 versus 26 retained items), and an observer-based affective measure that served only as corroborating evidence. Moreover, because the Case A ecological-awareness instrument was knowledge-referenced, its dimension-level gain pattern cannot be attributed solely to the medium. Finally, the study's three motivating gaps were derived from published systematic reviews and meta-analyses rather than a new systematic search of primary studies, so the gap analysis inherits the coverage and selection limits of those syntheses. Future research should test the propositions using controlled designs, larger and more diverse samples, delayed posttests, and cases that populate the remaining interactivity-by-outcome cells, particularly high-interactivity AR targeting affective outcomes and low-interactivity AR targeting cognitive outcomes, to provide a more direct test of Proposition 3.

CONCLUSION

Developed with ADDIE and evaluated as two development cases, AR eco-visualization media for secondary-science environmental topics met feasibility criteria based on expert agreement, were rated practical by teachers and students, and showed medium-sized pretest-posttest gains in both affective and cognitive outcomes. Because each case is a single development cycle without a control group, and because the two cases differ simultaneously in interactivity, outcome type, grade, school, and topic, these results characterize what each medium achieved in its own setting; they do not isolate interactivity as the cause of the difference between them. On RQ1 and RQ2, the answer is affirmative but bounded: both media were feasible and practical, and both showed medium-sized gains. On RQ3, the answer is a framework rather than a finding. The study's contribution is accordingly not the individual media, nor a demonstrated interactivity-outcome relationship, but a cross-case design framework offered for testing. It comprises six propositions linking the interactivity an AR medium affords to the type of learning goal it serves, grounded in ICAP and multimedia-learning accounts, informed by evidence on the internal structure of the gains, and developed under realistic, low-infrastructure conditions. For researchers, the framework offers hypotheses that a controlled design could falsify; Proposition 3 is the most direct target, and testing it requires the two off-diagonal cells this study did not populate. For science teachers and instructional designers, it offers a provisional allocation rule. For awareness-oriented instruction, recognition-based AR may be considered when accompanied by explicit reflection and action tasks. For conceptual understanding, manipulation-oriented AR may be considered when embedded within structured pedagogical support such as flipped pre-work. These implications remain provisional and require testing across matched learning goals and interactivity conditions.

AUTHOR CONTRIBUTIONS

MNH contributed to conceptualization, methodology development, formal analysis, funding acquisition, project administration, supervision, original drafting, and manuscript review. RMP contributed to conceptualization, supervision, resource provision, and manuscript review. KDHG developed the methodology and validated and edited the manuscript. NDF was involved in investigation, data collection, data curation, and manuscript review. DN was involved in investigation, data collection, data curation, and manuscript review. FCW validated, visualized, and edited the manuscript.

ACKNOWLEDGEMENTS

This research was funded by the Universitas Sebelas Maret Budget for the 2026 Fiscal Year through the FUNDAMENTAL RESEARCH A (PFA-UNS) Research scheme, Research Assignment Agreement Number 460/UN27.22/PT.01.03/2026.

REFERENCES

- Akçayır, M., & Akçayır, G. (2017). Advantages and challenges associated with augmented reality for education: A systematic review of the literature. *Educational Research Review*, 20, 1–11. <https://doi.org/10.1016/j.edurev.2016.11.002>
- Alalwan, N., Cheng, L., Al-Samarraie, H., Yousef, R., Alzahrani, A. I., & Sarsam, S. M. (2020). Challenges and prospects of virtual reality and augmented reality utilization among primary school teachers: A developing country perspective. *Studies in Educational Evaluation*, 66, 100876. <https://doi.org/10.1016/j.stueduc.2020.100876>
- AlGerafi, M. A. M., Zhou, Y., Oubibi, M., & Wijaya, T. T. (2023). Unlocking the potential: A comprehensive evaluation of augmented reality and virtual reality in education. *Electronics*, 12(18), 3953. <https://doi.org/10.3390/electronics12183953>
- Anesha, P., Irwandani, & Solviana, M. D. (2026). Systematic literature review: Efektivitas penggunaan media pembelajaran berbasis augmented reality pada pembelajaran biologi. *Spizaetus: Jurnal Biologi Dan Pendidikan Biologi*, 7(2), 199–212. <https://doi.org/10.55241/spibio.v7i2.713>
- Annan-Diab, F., & Molinari, C. (2017). Interdisciplinarity: Practical approach to advancing education for sustainability and for the Sustainable Development Goals. *The International Journal of Management Education*, 15(2), 73–83. <https://doi.org/10.1016/j.ijme.2017.03.006>
- Attanasi, G., Buljat Raymond, B., Festré, A., & Guido, A. (2025). Raising environmental awareness with augmented reality. *Ecological Economics*, 233, 108563. <https://doi.org/10.1016/j.ecolecon.2025.108563>
- Aviyanti, L., Fratiwi, N. J., Nurdini, N., Salam, A., & Nawas, A. (2025). Can multiple-choice items measure critical thinking in socio-scientific environmental issues? Evidence from a global warming assessment of grade 10 students using rasch analysis. *Jurnal Ilmiah Pendidikan Fisika Al-Biruni*, 14(2), 33–53. <https://doi.org/10.24042/jipfalbiruni.v14i2.28194>
- Barcenas, S. M., & Prudente, M. (2026). The effect of integrating augmented reality on students' cognitive learning outcomes in science education: A meta-analysis of recent literature. *Research and Practice in Technology Enhanced Learning*, 22, 018. <https://doi.org/10.58459/rptel.2027.22018>
- Branch, R. M. (2009). *Instructional design: The ADDIE approach*. Springer US. <https://doi.org/10.1007/978-0-387-09506-6>
- Castillo-Gonzalez, W., Lepez, C. O., & Bonardi, M. C. (2023). Augmented reality and environmental education: Strategy for greater awareness. *Gamification and Augmented Reality*, 1, 10. <https://doi.org/10.56294/gr202310>
- Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4), 219–243. <https://doi.org/10.1080/00461520.2014.965823>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- de Jong, T., & van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research*, 68(2), 179–201. <https://doi.org/10.3102/00346543068002179>
- Debrah, J. K., Vidal, D. G., & Dinis, M. A. P. (2021). Raising awareness on solid waste management through formal education for sustainability: A developing countries evidence review. *Recycling*, 6(1), 6. <https://doi.org/10.3390/recycling6010006>
- Febriyanti, N. D., Hudha, M. N., Probosari, R. M., Khasanah, A. N., & Supurwoko, S. (2026). Development of an integrated PhET and augmented reality package in a flipped classroom to improve conceptual understanding of renewable energy. *Jurnal Pembelajaran, Bimbingan, dan Pengelolaan Pendidikan*, 6(12), 1. <https://doi.org/10.17977/um065.v6.i12.2026.1>
- Fujii, S. (2006). Environmental concern, attitude toward frugality, and ease of behavior as determinants of pro-environmental behavior intentions. *Journal of Environmental Psychology*, 26(4), 262–268. <https://doi.org/10.1016/j.jenvp.2006.09.003>
- Furqon, M. (2024). The impact of utilizing interactive conceptual instruction assisted by PhET simulations on students' understanding ability in physics. *SPEKTRA: Jurnal Kajian Pendidikan Sains*, 10(2), 244–258. <https://doi.org/10.32699/spektra.v10i2.7740>

- Garzón, J., Pavón, J., & Baldiris, S. (2019). Systematic review and meta-analysis of augmented reality in educational settings. *Virtual Reality*, 23(4), 447–459. <https://doi.org/10.1007/s10055-019-00379-9>
- Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66(1), 64–74. <https://doi.org/10.1119/1.18809>
- Hamilton, D., McKechnie, J., Edgerton, E., & Wilson, C. (2020). Immersive virtual reality as a pedagogical tool in education: A systematic literature review of quantitative learning outcomes and experimental design. *Journal of Computers in Education*, 8(1), 1–32. <https://doi.org/10.1007/s40692-020-00169-2>
- Handayani, J. A., & Widodo, A. (2024). The impact of student participation in environmental education programs on pro-environmental behavior. *Jurnal Cakrawala Pendas*, 10(2), 215–223. <https://doi.org/10.31949/jcp.v10i2.8475>
- Handhika, J., Alvarez, J. I., Pramono, N. A., & Mayasari, T. (2025). Profile of Student Conceptions on the Direct Current Circuits with multiple-activity-representation. *Jurnal Ilmiah Pendidikan Fisika Al-Biruni*, 14(2), 167–186. <https://doi.org/10.24042/jipfalbiruni.v14i2.29152>
- Harding, D., Kadiyono, A. L., Hafiar, H., & Wibowo, H. (2018). Mind the gap: What are the barriers to pro-environmental behavior among students? *Journal of Business and Social Review in Emerging Economies*, 4(1), 1–6. <https://doi.org/10.26710/jbsee.v4i1.351>
- Hu, B., Jia, W., Mi, S., & Bi, H. (2026). The impact of computer simulation on students' conceptual understanding in K-12 science education: A meta-analysis. *Journal of Science Education and Technology*, 35(2), 534–553. <https://doi.org/10.1007/s10956-025-10264-7>
- Huda, C., Hudha, M. N., Ain, N., Nandiyanto, A. B. D., Abdullah, A. G., & Widiaty, I. (2018). The implementation of blended learning using android-based tutorial video in computer programming course II. *IOP Conference Series: Materials Science and Engineering*, 288, 012163. <https://doi.org/10.1088/1757-899x/288/1/012163>
- Hudha, M. N., Batlolona, J. R., & Wartono. (2019). *Science literacy ability and physics concept understanding in the topic of work and energy with inquiry-STEM*. 2202, 020063. <https://doi.org/10.1063/1.5141676>
- Hudha, M. N., Hamidah, I., Permanasari, A., & Abdullah, A. G. (2021). How low-carbon issues are addressed in primary school textbooks. *Jurnal Pendidikan IPA Indonesia*, 10(2), 260–269. <https://doi.org/10.15294/jpii.v10i2.26628>
- Kalemkuş, J., & Kalemkuş, F. (2022). Effect of the use of augmented reality applications on academic achievement of student in science education: Meta-analysis review. *Interactive Learning Environments*, 31(9), 6017–6034. <https://doi.org/10.1080/10494820.2022.2027458>
- Keller, S., Rumann, S., & Habig, S. (2021). Cognitive load implications for augmented reality supported chemistry learning. *Information*, 12(3), 96. <https://doi.org/10.3390/info12030096>
- Khan, E. D., Khan, D. M., & Malik, H. W. (2026). Harnessing artificial intelligence for sustainable education: Integrating environmental awareness and SDG alignment in learning systems. *Journal of Environment and Sustainability Education*, 4(1), 192–204. <https://doi.org/10.62672/joease.v4i1.239>
- Khofifah, K., Yuliani, H., & Santiani, S. (2024). Meta-analysis: The effect of PhET simulation media on enhancing conceptual understanding in physics learning. *Jurnal Ilmiah Pendidikan Fisika*, 7(3), 532–542. <https://doi.org/10.20527/jipf.v7i3.9046>
- Kioupi, V., & Voulvoulis, N. (2019). Education for sustainable development: A systemic framework for connecting the SDGs to educational outcomes. *Sustainability*, 11(21), 6104. <https://doi.org/10.3390/su11216104>
- Kollmuss, A., & Agyeman, J. (2002). Mind the gap: Why do people act environmentally and what are the barriers to pro-environmental behavior? *Environmental Education Research*, 8(3), 239–260. <https://doi.org/10.1080/13504620220145401>
- Küçük Avcı, Ş., Çoklar, A., & İstanbullu, A. (2019). The effect of three-dimensional virtual environments and augmented reality applications on the learning achievement: A meta-analysis study. *Education and Science*, 44(198), 149–182. <https://doi.org/10.15390/eb.2019.7969>

- Laaloua, H. (2026). Environmental literacy in Morocco: Uncovering the knowledge/attitude gap in environmental education. *International Journal of Social Science and Human Research*, 09(06). <https://doi.org/10.47191/ijsshr/v9-i6-56>
- Li, F., Wang, X., He, X., Cheng, L., & Wang, Y. (2021). How augmented reality affected academic achievement in K-12 education – A meta-analysis and thematic-analysis. *Interactive Learning Environments*, 31(9), 5582–5600. <https://doi.org/10.1080/10494820.2021.2012810>
- Marougkas, A., Troussas, C., Krouska, A., & Sgouropoulou, C. (2023). Virtual reality in education: A review of learning theories, approaches and methodologies for the last decade. *Electronics*, 12(13), 2832. <https://doi.org/10.3390/electronics12132832>
- Mayer, R. E. (2014). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (2nd ed., pp. 43–71). Cambridge University Press. <https://doi.org/10.1017/CBO9781139547369.005>
- Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). SAGE Publications.
- Nugroho, D. A., Wahyu, W., & Casmana, A. R. (2025). Engaging learning for environmental awareness: Integrating virtual reality in ecological citizenship education. *Journal of Moral and Civic Education*, 9(1), 36–47. <https://doi.org/10.24036/8851412912025865>
- Ozturk, E., & Akcay, G. (2023). Can environmental education supported by augmented reality (AR) applications improve the environmental awareness of primary school students? *The Eurasia Proceedings of Educational and Social Sciences*, 31, 216–229. <https://doi.org/10.55549/epess.1381983>
- Pranata, O. D. (2023). Enhancing conceptual understanding and concept acquisition of gravitational force through guided inquiry utilizing PhET simulation. *Sainstek: Jurnal Sains Dan Teknologi*, 15(1), 44–53. <https://doi.org/10.31958/js.v15i1.9191>
- Rahmawati, Y., Zulhipri, Hartanto, O., Falani, I., & Iriyadi, D. (2022). Students' conceptual understanding in chemistry learning using PhET interactive simulations. *Journal of Technology and Science Education*, 12(2), 303–326. <https://doi.org/10.3926/jotse.1597>
- Santos, M. E. C., Chen, A., Taketomi, T., Yamamoto, G., Miyazaki, J., & Kato, H. (2014). Augmented reality learning experiences: Survey of prototype design and evaluation. *IEEE Transactions on Learning Technologies*, 7(1), 38–56. <https://doi.org/10.1109/tlt.2013.37>
- Saputra, A., Wilujeng, I., Wiyarsi, A., & Niyamae, A. N. (2026). Augmented reality in science education: Bibliometric and systematic literature review of trends and outcomes. *Research and Development in Education (RaDEn)*, 6(1), 105–137. <https://doi.org/10.22219/raden.v6i1.43412>
- Suherman, W., Kariadinata, R., & Chusni, M. M. (2025). Effectiveness of PhET interactive simulation-based learning on Grade XI students' conceptual understanding of projectile motion. *CAHAYA: Journal of Research on Science Education*, 3(2), 57–71. <https://doi.org/10.70115/cahaya.v3i2.331>
- Suo, X., Yin, B., & Feng, X. (2026). Augmented reality and embodied learning: Effects of embodiment degrees on students' learning achievement, cognitive load, and technology acceptance. *Frontiers in Psychology*, 16, 1712261. <https://doi.org/10.3389/fpsyg.2025.1712261>
- Suyidno, S., Misbah, M., Mastuang, M., Komariyah, L., Qamariah, Q., Amiruddin, M. Z. Bin, & Rahman, N. F. A. (2025). Evolution of simulation-driven science education: Three decades of scholarly patterns and research dynamics (1996–2025). *Online Learning in Educational Research*, 6(1), 29–47. <https://doi.org/10.58524/oler.v6i1.1000>
- Suzuki, Y., Wild, F., & Scanlon, E. (2023). Measuring cognitive load in augmented reality with physiological methods: A systematic review. *Journal of Computer Assisted Learning*, 40(2), 375–393. <https://doi.org/10.1111/jcal.12882>
- Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, 31(2), 261–292. <https://doi.org/10.1007/s10648-019-09465-5>
- Taufik, L. M., Widodo, A., Surtikanti, H. K., & Rahman, T. (2026). Development of a systems-thinking-based assessment instrument for environmental literacy and problem-solving skills in SDG 6 contexts. *Journal of Environment and Sustainability Education*, 4(1), 1–11. <https://doi.org/10.62672/joease.v4i1.173>

- Tsai, Y.-L., & Tsai, C.-C. (2020). A meta-analysis of research on digital game-based science learning. *Journal of Computer Assisted Learning*, 36(3), 280–294. <https://doi.org/10.1111/jcal.12430>
- Tzima, S., Styliaras, G., & Bassounas, A. (2019). Augmented reality applications in education: Teacher's point of view. *Education Sciences*, 9(2), 99. <https://doi.org/10.3390/educsci9020099>
- Van Alten, D. C. D., Phielix, C., Janssen, J., & Kester, L. (2019). Effects of flipping the classroom on learning outcomes and satisfaction: A meta-analysis. *Educational Research Review*, 28, 100281. <https://doi.org/10.1016/j.edurev.2019.05.003>
- Wang, Q., & Li, Y. (2024). How virtual reality, augmented reality and mixed reality facilitate teacher education: A systematic review. *Journal of Computer Assisted Learning*, 40(3), 1276–1294. <https://doi.org/10.1111/jcal.12949>
- Wu, H.-K., Lee, S. W.-Y., Chang, H.-Y., & Liang, J.-C. (2013). Current status, opportunities and challenges of augmented reality in education. *Computers & Education*, 62, 41–49. <https://doi.org/10.1016/j.compedu.2012.10.024>
- Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). SAGE Publications.
- Yu, Q., & Yu, K. (2024). The effects of gamified flipped classroom on student learning: Evidence from a meta-analysis. *Interactive Learning Environments*, 32(9), 5126–5141. <https://doi.org/10.1080/10494820.2023.2209791>
- Zhu, X., Peng, K., Yu, S., & Wang, G. (2026). Can augmented reality technology reduce learners' cognitive load? A meta-analysis. *Smart Learning Environments*, 13(1). <https://doi.org/10.1186/s40561-025-00429-7>