

## Interactive flipbook-based e-modules and fifth-grade students' learning outcomes in the properties of light: a quasi-experimental study with baseline adjustment

Reni Nur Habibah, Widi Wulansari, Encil Puspitonigrum

Program Studi Pendidikan Guru Sekolah Dasar, Universitas Nusantara PGRI Kediri, Indonesia

\*) Correspondence regarding this article should be addressed to: Author address e-mail: renihabibah606@gmail.com

**Abstract:** Properties of light are difficult for elementary students because the underlying processes can only be observed through their effects, and flipbook-based e-modules are often promoted as a way to make these processes visible. Most evidence for such e-modules, however, comes from development studies and single-group pre-test-post-test comparisons. This study tested whether an interactive flipbook-based e-module produced higher learning outcomes on the properties of light than regular instruction once pre-existing achievement was taken into account. A quasi-experimental pre-test-post-test control-group design was used with two intact fifth-grade classes at a public elementary school in Indonesia ( $N = 22$ ; experimental  $n = 12$ , control  $n = 10$ ). Both classes took the same 0–100 achievement test before and after instruction. The experimental class started 11.92 points higher (Hedges  $g = 0.73$ ), so the primary analysis was an analysis of covariance (ANCOVA) with the pre-test as covariate, supported by permutation, bootstrap, rank-based, and robust sensitivity analyses. Both classes improved substantially, by about 33 and 35 points. After baseline adjustment, the e-module class did not outperform the control class,  $F(1, 19) = 0.25$ ,  $p = .626$ , partial  $\eta^2 = .013$ ; adjusted difference =  $-3.38$  points, 95% CI  $[-17.65, 10.90]$ , and all sensitivity analyses agreed. The e-module was therefore not superior to regular instruction in this setting, although the small sample cannot exclude moderate effects in either direction. The study also shows that within-group gains and normalised gain can overstate the benefit of digital learning materials when baseline differences between intact classes are ignored.

**Keywords:** elementary science education; interactive e-module; learning outcomes; properties of light; quasi-experimental design.

**Article History:** Received on 28/08/2026; Revised on 15/10/2026; Accepted on 16/10/2026; Published Online: 17/10/2026.



This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. ©2026 by author.

## INTRODUCTION

Elementary science education aims at more than the recall of facts. National curricula, including Indonesia's Merdeka Curriculum for the integrated Natural and Social Sciences subject (Ilmu Pengetahuan Alam dan Sosial, IPAS), expect students to connect observations with explanations and to build simple scientific models of everyday phenomena. Reaching this goal depends partly on

whether children can picture processes that cannot be seen directly, and partly on whether learning resources help them link those processes to familiar experiences.

Light is a clear example of such content. Its propagation, reflection, and refraction are known only through their effects, such as illuminated objects, images in mirrors, and shadows. Research on alternative conceptions has long shown that learners treat light as something that fills a space rather than something that travels from a source, or confuse light with its effects, such as brightness and shadow (Fetherstonhaugh & Treagust, 1992; Bendall et al., 1993). These ideas persist into adulthood. A recent survey of 132 prospective elementary teachers in Quebec found that most could not explain everyday cases of reflection, refraction, absorption, and shadow formation; many believed, for example, that light stops when it reaches white paper or that the glass of a mirror is what reflects light (Métoui, 2023). When future teachers hold such views, their pupils are likely to meet them as well, and explanation alone rarely changes them (Fetherstonhaugh & Treagust, 1992).

A common teaching response is to make invisible processes visible. Ray diagrams, photographs, animations, and manipulable displays can show the path of light, the law of reflection, or the bending of light at a boundary. The cognitive theory of multimedia learning offers a rationale for this practice: learners process verbal and pictorial information through separate channels, and well-designed combinations of words and pictures help them select relevant information, organise it, and integrate it with prior knowledge, provided that the design avoids unnecessary processing (Mayer & Moreno, 1998; Mayer, 2024). Cognitive load theory adds that such benefits depend on whether a format reduces extraneous load compared with the alternative (Sweller, 1988; Sweller et al., 2019). An overview of meta-analyses on multimedia design confirmed that several design principles improve learning, while showing that the size of the benefit varies with the comparison being made (Noetel et al., 2022).

Digital learning materials have spread quickly in elementary classrooms as access to devices has grown. Unlike printed textbooks, they can combine text, images, video, navigation, and practice items in one resource. Students tend to view visualization technologies such as three-dimensional models and augmented reality favourably (Abdinejad et al., 2021), and the educational use of these technologies is expanding (Al-Ansi et al., 2023). Meta-analytic evidence is more cautious. In a context-specific meta-analysis of mathematics and science learning, Hillmayr et al. (2020) found an overall positive effect of digital tools, but the effect depended on the type of tool, on teacher training, and on whether the tools supplemented or replaced other forms of instruction.

Interactive electronic modules (e-modules) are digital learning packages designed for relatively independent use. They usually contain explanations, worked examples, illustrations, exercises, and self-assessment, and they allow non-linear navigation. In elementary science, e-modules have been linked to higher motivation and to favourable usability ratings from teachers and students (Lase & Sudarma, 2023; Alyusfitri et al., 2024). The flipbook format is a popular way to deliver them: authoring tools convert a document into a page-turning publication that can embed images, audio, video, and quizzes. Indonesian researchers have produced many flipbook e-modules and have reported high expert-validity and practicality scores, often together with gains from pre-test to post-test (Winatha, 2018; Widiana & Rosy, 2021; Munzil et al., 2022; Rohmatin et al., 2022; Sabil et al., 2023).

This body of evidence is encouraging but methodologically uneven. Much of it comes from research-and-development studies that judge a module by expert validation or by comparing the pre-test and post-test scores of a single group (Mahmudah et al., 2022; Saralee et al., 2024). Such studies show that a module can be built and used and that scores rise after teaching, yet a single-group gain cannot separate the effect of the module from maturation, familiarity with the test, or the ordinary instruction that accompanies it (Shadish et al., 2002). Studies with a control class do exist,

but they usually rely on post-test comparisons or normalised gain (Maulidya & Zulherman, 2024; Arina et al., 2025). Normalised gain depends on the pre-test level and can rank groups differently from standardised effect sizes (Hake, 1998; Nissen et al., 2018), and a post-test comparison of intact classes may attribute a pre-existing difference to the intervention. Analysing post-test scores with the pre-test as a covariate adjusts for such imbalance and is more precise than analysing change scores (Vickers & Altman, 2001).

Two gaps follow from this review. First, controlled studies of flipbook e-modules in elementary science that explicitly adjust for baseline achievement, and that report effect sizes with confidence intervals, remain scarce. Without them it is difficult to tell how much of a reported advantage reflects the module and how much reflects the classes being compared. Second, the properties of light, a topic where documented alternative conceptions make visual support especially plausible, have received less attention than broad thematic content. The present study addresses both gaps. Its contribution is methodological as well as substantive: it applies a baseline-adjusted analysis with several sensitivity checks to a type of intervention that is usually evaluated through within-group gains, and it reports the result regardless of its direction.

The study therefore asked the following question: among fifth-grade students learning the properties of light, does an interactive flipbook-based e-module produce higher post-test achievement than regular instruction after pre-test achievement is controlled? The independent variable was the type of instruction (e-module versus regular instruction), and the dependent variable was the score on a 0–100 achievement test on the properties of light. In line with multimedia learning theory, we hypothesised that the e-module class would obtain higher adjusted post-test scores. The analysis treated a null or reversed result as equally informative. The baseline-adjusted difference between groups, rather than the change within each group, served as the primary evidence, and effect sizes with 95% confidence intervals were reported so that the magnitude and precision of any effect could be judged alongside statistical significance.

## METHOD

### Research Design

The study used a quasi-experimental pre-test–post-test control-group design, also known as a non-equivalent groups design, with two fifth-grade classes (Shadish et al., 2002). The design is quasi-experimental because students were not randomly assigned to conditions. The existing classes were kept intact and served as the units of assignment, a common arrangement in schools where regrouping pupils for research would disrupt teaching. The main threat to internal validity in this design is a pre-existing difference between classes. It was handled in three ways: baseline equivalence was tested and its magnitude reported, the primary analysis adjusted post-test scores for pre-test scores, and the robustness of the adjusted estimate was examined with four sensitivity analyses. Figure 1 summarises the design, timeline, and analysis plan.

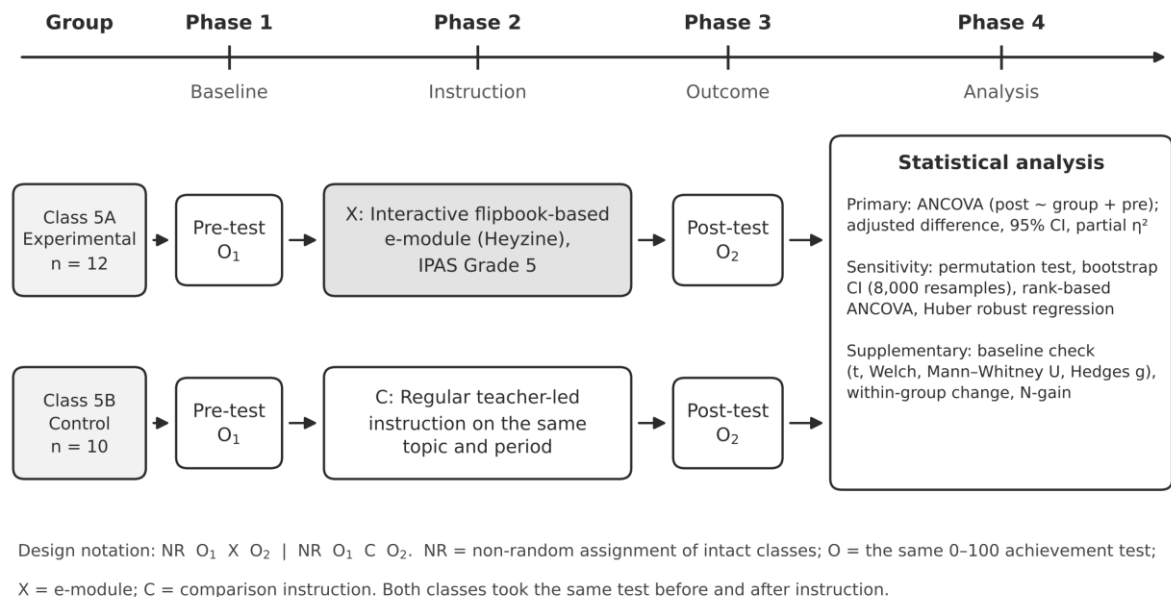


**Figure 1.** Research design, timeline, and analysis plan. Two intact fifth-grade classes took the same achievement test before (O<sub>1</sub>) and after (O<sub>2</sub>) instruction on the properties of light; the experimental class used the interactive flipbook-based e-module (X), and the control class received regular instruction (C). Classes, not individual students, were assigned to conditions.

### Participants and Setting

Participants were 22 fifth-grade students from two intact classes at SDN Grogol 3, a public elementary school in Indonesia that teaches IPAS under the Merdeka Curriculum. Class 5A formed the experimental group (n = 12; 7 boys and 5 girls), and Class 5B formed the control group (n = 10; 5 boys and 5 girls). All students were 11 years old and attended every session (100% attendance), so age and attendance did not differ between groups. Student names were replaced with codes (A01–A12 and B01–B10) before analysis. The study was carried out with the permission and cooperation of the school principal and the class teachers, and no identifying information was retained in the data set.

### Sampling Procedure

Two complete fifth-grade classes were selected, and every student in them was included (total sampling within the selected classes). The analytic sample consisted of all students who took part in the instruction and completed both tests; no student was excluded and no data were missing. Because classes rather than individuals were assigned to conditions, and the classes were not drawn at random, the sample is a non-probability sample and cannot be treated as random, stratified, or cluster sampling. The findings should therefore be generalised with caution.

### Interactive Flipbook-Based E-Module

The experimental class learned the properties of light with an interactive flipbook-based e-module created with the Heyzine flipbook maker and published as a web-based, page-turning book (<https://heyzine.com/flip-book/ac1653b62d.html>). The module was written for the Grade 5 IPAS subject of the Merdeka Curriculum. It contained a preface, instructions for use, a navigation guide, the curriculum learning outcomes and objectives, an embedded learning video, the core material, a student worksheet (Lembar Kerja Peserta Didik, LKPD), an evaluation section, a list of references, and a profile of the developer.

The material covered six properties of light: rectilinear propagation (light travels in straight lines), reflection, transmission through transparent objects, refraction, dispersion of white light into its spectral colours, and the formation of shadows when light is blocked by opaque objects. Each property was explained in text alongside a picture of an everyday situation. Interactive features included page navigation, zoom and display modes, audio clips, the embedded video, and online activities and quizzes that required an internet connection. The learning objectives required students to identify the properties of light (C1), analyse the relationship between the properties of light and the way we see (C4), and demonstrate the properties of light through a simple practical activity (P3).

### Control Condition

The control class studied the same topic during the same period through the school's regular teacher-led instruction, without access to the e-module. Instruction in the control class followed the school's usual practice for the topic. The content, number, and duration of the control lessons were not documented in the same detail as the e-module, so the comparison is between the e-module and usual classroom practice rather than a fully specified alternative method; this is acknowledged as a limitation.

### Learning-Outcome Instrument

Learning outcomes were measured with an achievement test on the properties of light, scored from 0 to 100. The same test served as pre-test and post-test so that both scores were on the same scale. The test covered the properties of light taught in both classes and was administered under the same conditions to both groups. Formal evidence of the validity and reliability of the test (such as expert content validation or an internal-consistency coefficient) was not available for this study; the scores are therefore treated as a classroom achievement measure, and this is considered when interpreting the results. Raw gain (post-test minus pre-test) and normalised gain,  $(\text{post-test} - \text{pre-test}) / (100 - \text{pre-test})$ , were computed from these scores.

### Intervention Procedure

The study followed three stages (Figure 1). Both classes first took the pre-test before any instruction on the properties of light. The experimental class then studied the topic with the flipbook e-module, while the control class received regular instruction on the same topic. After instruction ended, both classes took the post-test.

### Data Analysis

Analyses followed five stages. First, descriptive statistics (mean, standard deviation, median, range, and 95% confidence interval of the mean) were calculated for the pre-test, post-test, raw gain, and normalised gain of each group. Second, baseline comparability was examined by comparing pre-test scores with an independent-samples t-test, Welch's t-test, and the Mann-Whitney U test. Hedges' *g* with its 95% CI served as the standardised effect size, and the magnitude of the imbalance was given more weight than its *p*-value, because a non-significant test in a small sample does not establish equivalence (Lakens, 2017). Third, distributional assumptions were inspected with histograms, boxplots, and individual data points, and with the Shapiro-Wilk and Levene tests, which were interpreted cautiously given the sample size.

Fourth, the primary treatment effect was estimated with analysis of covariance (ANCOVA), modelling post-test scores as a function of group (experimental = 1, control = 0) with the pre-test score as covariate:  $\text{Post} = \beta_0 + \beta_1 \cdot \text{Group} + \beta_2 \cdot \text{Pre} + \varepsilon$ . The coefficient  $\beta_1$  is the baseline-adjusted difference between groups. ANCOVA was preferred to a comparison of raw gains because it adjusts for baseline differences and estimates the group effect more precisely (Vickers & Altman, 2001). The assumption of homogeneous regression slopes was tested by adding a group  $\times$  pre-test interaction. The adjusted difference was reported with its 95% CI and partial  $\eta^2$ , and adjusted means were evaluated at the grand-mean pre-test score. Within-group change was assessed with paired t-tests,

Wilcoxon signed-rank tests, and the standardised mean change ( $d_z$ ) for completeness only, because change within a group does not show that one condition is better than another.

Fifth, the adjusted effect was checked with four sensitivity analyses: a permutation test of the ANCOVA group coefficient, a bootstrap CI for the adjusted difference (8,000 resamples), a rank-based ANCOVA, and a robust (Huber) regression. Normalised gain was reported only as supplementary descriptive evidence because it is sensitive to the pre-test level (Hake, 1998; Nissen et al., 2018). Effect sizes were interpreted with reference to Cohen (1988) and to benchmarks for educational interventions, which indicate that effects on achievement tests are usually much smaller than Cohen's conventions suggest (Kraft, 2020). Because the sample size was fixed by the two available classes, no a priori power analysis was possible; instead, the minimum detectable effect of the design was calculated from the standard error of the adjusted difference. The significance level was  $\alpha = .05$ . Analyses were conducted in Python 3 with NumPy, pandas, SciPy, and statsmodels, and results are reported following the conventions of the APA Publication Manual, 7th edition (American Psychological Association, 2020).

## RESULTS AND DISCUSSION

### Participant and Data Characteristics

All 22 students were included in the analysis, with no missing pre-test or post-test scores and no duplicate records. Raw and normalised gains recalculated independently from the pre-test and post-test scores matched the values in the study records exactly, and no impossible or out-of-range scores were found. Age (11 years) and attendance (100%) were constant across the sample. Table 1 summarises the characteristics of the participants.

**Table 1.** Participant Characteristics by Study Group

| Characteristic | Experimental | Control | Total |
|----------------|--------------|---------|-------|
| n              | 12           | 10      | 22    |
| Boys, n        | 7            | 5       | 12    |
| Girls, n       | 5            | 5       | 10    |
| Age (years)    | 11           | 11      | 11    |
| Attendance (%) | 100          | 100     | 100   |

*Note.* Age and attendance were constant across all participants. The experimental group used the interactive flipbook-based e-module; the control group received comparison instruction.

### Baseline Learning Outcomes

Table 2 presents descriptive statistics for all outcome variables. Before instruction, the experimental class scored higher on average than the control class ( $M = 50.42$ ,  $SD = 14.22$  vs.  $M = 38.50$ ,  $SD = 17.33$ ), a difference of 11.92 points. The difference was not statistically significant,  $t(20) = 1.77$ ,  $p = .091$  (Welch  $t = 1.74$ ,  $p = .099$ ; Mann-Whitney  $U = 86.0$ ,  $p = .090$ ), but its standardised size was medium to large, Hedges  $g = 0.73$ , 95% CI  $[-0.11, 1.57]$  (Table 3). The interval extends from a negligible to a very large advantage for the experimental class. With  $N = 22$ , the non-significant result is as likely to reflect low power as genuine similarity, so baseline equivalence cannot be assumed (Lakens, 2017). This imbalance is the reason the primary analysis adjusted for the pre-test. Figure 2 shows the score distributions of both groups at pre-test and post-test.

**Table 2.** Descriptive Statistics for Learning Outcomes by Study Group

| Variable                     | n  | M     | SD    | Mdn   | Range  | 95% CI for M   |
|------------------------------|----|-------|-------|-------|--------|----------------|
| <i>Experimental (n = 12)</i> |    |       |       |       |        |                |
| Pre-test                     | 12 | 50.42 | 14.22 | 52.50 | 25–70  | [41.38, 59.45] |
| Post-test                    | 12 | 83.33 | 21.25 | 90.00 | 30–100 | [69.83, 96.83] |
| Raw gain                     | 12 | 32.92 | 13.05 | 35.00 | 5–55   | [24.63, 41.21] |

| Variable                | n  | M     | SD    | Mdn   | Range     | 95% CI for M   |
|-------------------------|----|-------|-------|-------|-----------|----------------|
| N-gain                  | 12 | 0.71  | 0.29  | 0.76  | 0.07–1.00 | [0.53, 0.90]   |
| <b>Control (n = 10)</b> |    |       |       |       |           |                |
| Pre-test                | 10 | 38.50 | 17.33 | 45.00 | 5–55      | [26.10, 50.90] |
| Post-test               | 10 | 73.50 | 24.16 | 77.50 | 40–100    | [56.22, 90.78] |
| Raw gain                | 10 | 35.00 | 16.16 | 37.50 | 10–55     | [23.44, 46.56] |
| N-gain                  | 10 | 0.61  | 0.33  | 0.55  | 0.15–1.00 | [0.37, 0.85]   |

Note. M = mean; SD = standard deviation; Mdn = median; CI = confidence interval. Raw gain = post-test – pre-test. N-gain = (post-test – pre-test)/(100 – pre-test). Scores range 0–100.

**Table 3.** Baseline (Pre-test) Comparison Between Groups

| Statistic          | Experimental  | Control       | Value                    |
|--------------------|---------------|---------------|--------------------------|
| Pre-test M (SD)    | 50.42 (14.22) | 38.50 (17.33) | -                        |
| Mean difference    |               |               | 11.92                    |
| Independent t-test |               |               | $t(20) = 1.77, p = .091$ |
| Welch t-test       |               |               | $t = 1.74, p = .099$     |
| Mann-Whitney U     |               |               | $U = 86.0, p = .090$     |
| Hedges g [95% CI]  |               |               | 0.73 [-0.11, 1.57]       |

Note. A non-significant test with  $N = 22$  does not establish equivalence; the medium-to-large  $g$  and wide CI indicate a non-trivial baseline imbalance, motivating covariate adjustment.

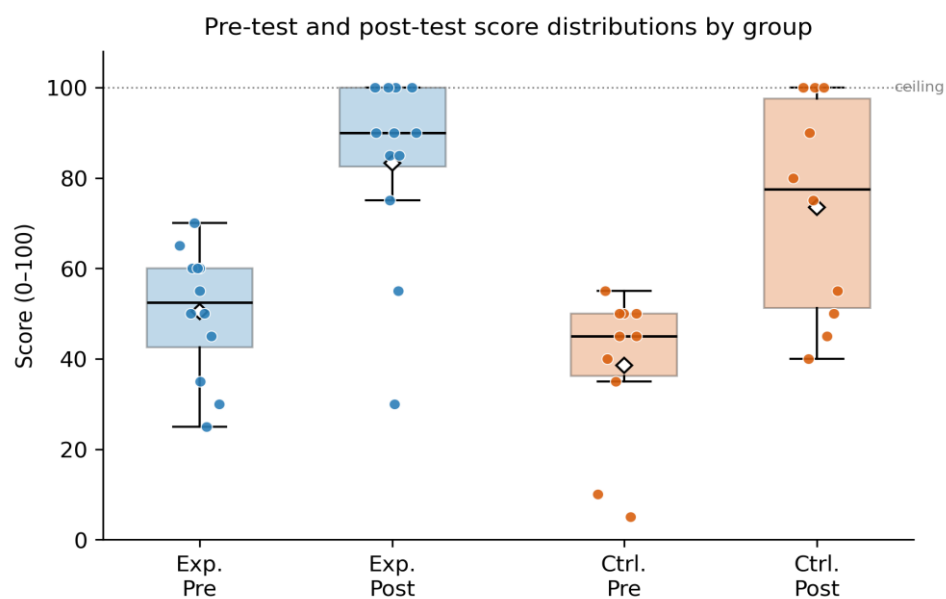


**Figure 2.** Pre-test and post-test score distributions by group. Boxes show the median and interquartile range, white diamonds show group means, and points show individual students (jittered horizontally). The dotted line marks the ceiling of the 0–100 scale. Several students in both groups reached the maximum score at post-test.

### Pre-test–Post-test Changes

Both groups improved markedly from pre-test to post-test (Table 4). Scores in the experimental group rose by 32.92 points on average (95% CI [24.63, 41.21]),  $t(11) = 8.74, p < .001, dz = 2.52$ , and the Wilcoxon signed-rank test agreed ( $p < .001$ ). Scores in the control group rose by 35.00 points (95% CI [23.44, 46.56]),  $t(9) = 6.85, p < .001, dz = 2.17$  (Wilcoxon  $p = .002$ ). Figure 3 shows the individual trajectories: most students in both classes improved substantially, while a few changed little. In raw points the average gain was slightly larger in the control group. Large gains therefore occurred under both conditions and cannot, on their own, be attributed to the e-module.

**Table 4.** Within-Group Pre-test–Post-test Changes

| Group        | Pre M | Post M | Change | Test                     | dz   |
|--------------|-------|--------|--------|--------------------------|------|
| Experimental | 50.42 | 83.33  | +32.92 | $t(11) = 8.74, p < .001$ | 2.52 |
| Control      | 38.50 | 73.50  | +35.00 | $t(9) = 6.85, p < .001$  | 2.17 |

Note. Wilcoxon signed-rank tests agreed (experimental  $p < .001$ ; control  $p = .002$ ). dz = standardised mean change. Within-group change does not by itself demonstrate between-group superiority.

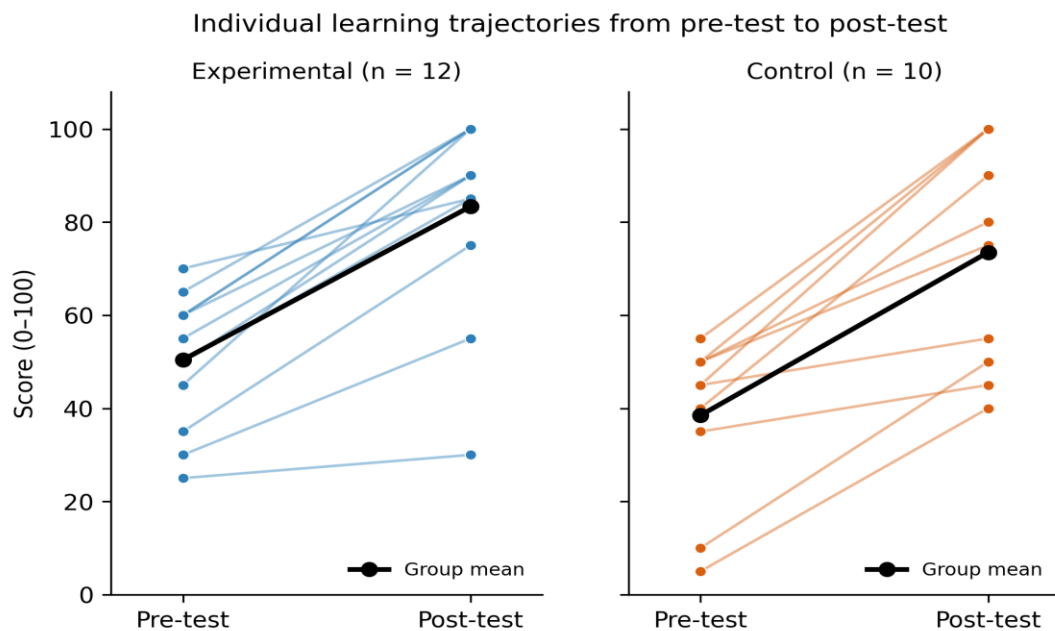


**Figure 3.** Individual learning trajectories from pre-test to post-test for the experimental (left) and control (right) groups. Thin coloured lines represent individual students; the heavy black line represents the group mean. Both groups improved on average, with visible variation among students in the amount of change.

### Primary Intervention Effect

In the primary ANCOVA, post-test scores were modelled with group as the factor and the pre-test as covariate. The group  $\times$  pre-test interaction was not significant ( $p = .719$ ), so the assumption of homogeneous regression slopes was retained. The pre-test was a strong predictor of the post-test ( $b = 1.11, p < .001$ ), consistent with the within-group pre–post correlations (experimental  $r = .80$ ; control  $r = .47$ ). After adjustment for the pre-test, the group effect was small and not significant,  $F(1, 19) = 0.25, p = .626$ , partial  $\eta^2 = .013$ . The adjusted post-test mean of the experimental group was slightly lower than that of the control group (77.33, 95% CI [68.07, 86.59] vs. 80.71, 95% CI [70.50, 90.92]), giving an adjusted difference of  $-3.38$  points, 95% CI  $[-17.65, 10.90]$  (Table 5; Figure 4). The interval includes zero and is wide enough to accommodate a moderate advantage for either group. The data therefore do not support the hypothesised benefit of the e-module, yet they cannot rule out a moderate benefit or a moderate disadvantage.

**Table 5.** ANCOVA of Post-test Scores (Covariate: Pre-test)

| Term                 | b     | SE   | 95% CI          | p      |
|----------------------|-------|------|-----------------|--------|
| Intercept            | 30.81 | 9.38 | [11.19, 50.44]  | < .001 |
| Group (Exp vs. Ctrl) | -3.38 | 6.82 | [-17.65, 10.90] | .626   |
| Pre-test (covariate) | 1.11  | 0.21 | [0.67, 1.55]    | < .001 |

Note.  $F(1, 19) = 0.25$  for group, partial  $\eta^2 = .013$ . Adjusted post-test means at grand-mean pre-test (45.0): experimental 77.33 [68.07, 86.59]; control 80.71 [70.50, 90.92]. Homogeneity-of-slopes interaction  $p = .72$ . Positive group coefficient would favour the experimental group; the estimate is negative and non-significant.

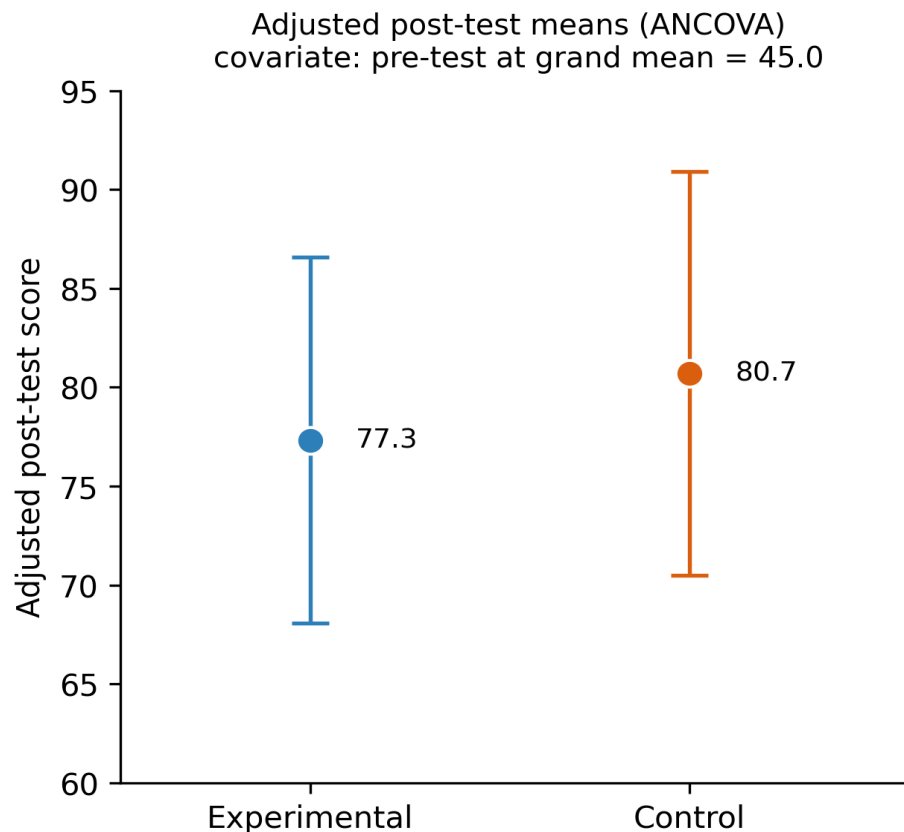


**Figure 4.** Adjusted post-test means from the ANCOVA, evaluated at the grand-mean pre-test score (45.0), with 95% confidence intervals. The adjusted means are statistically indistinguishable, and the interval for the experimental group lies slightly below that of the control group. These are model-adjusted estimates, not raw post-test means.

#### Gain, Normalised Gain, and Sensitivity Analyses

Supplementary analyses pointed in the same direction. Raw gains did not differ between groups (experimental  $M = 32.92$ ,  $SD = 13.05$ ; control  $M = 35.00$ ,  $SD = 16.16$ ),  $t(20) = -0.34$ ,  $p = .741$ , Hedges  $g = -0.14$ . Normalised gain was somewhat higher in the experimental group ( $M = 0.71$ ,  $SD = 0.29$ ) than in the control group ( $M = 0.61$ ,  $SD = 0.33$ ), but not significantly,  $t(20) = 0.79$ ,  $p = .437$ ,  $g = 0.33$ . This index is hard to interpret in the present data because, for the same raw gain, it gives higher values to students who start higher, which is precisely the situation here (Hake, 1998; Nissen et al., 2018). Using conventional thresholds, 8 of 12 experimental students and 4 of 10 control students reached the high normalised-gain category (Table 6), but these counts should not be over-interpreted given the small groups and the arbitrary cut-offs.

**Table 6.** Normalised-Gain Comparison and Category Distribution

| Index                   | Experimental | Control     | Between-group test          |
|-------------------------|--------------|-------------|-----------------------------|
| N-gain M (SD)           | 0.71 (0.29)  | 0.61 (0.33) | $t(20) = 0.79$ , $p = .437$ |
| High ( $\geq 0.70$ ), n | 8            | 4           | $g = 0.33$                  |
| Moderate (0.30–0.69), n | 3            | 4           | -                           |
| Low ( $< 0.30$ ), n     | 1            | 2           | -                           |

*Note.* Category thresholds follow Hake (1998). N-gain is mechanically sensitive to baseline differences and is reported as supplementary evidence only; category counts should not be over-interpreted given the small sample.

All four sensitivity analyses led to the same conclusion as the ANCOVA (Table 7). The permutation test of the adjusted group coefficient was not significant ( $p = .606$ ), the 95% bootstrap

CI of the adjusted difference included zero  $[-16.65, 12.14]$ , and neither the rank-based ANCOVA ( $p = .705$ ) nor the robust regression ( $p = .524$ ) showed a group effect. Because parametric, resampling, rank-based, and robust methods agree, the null result is not an artefact of a single modelling assumption. Figure 5 shows the distributions of normalised gain and the bootstrap distribution of the adjusted difference.

**Table 7.** Sensitivity Analyses of the Baseline-Adjusted Group Effect

| Method                    | Estimate / test     | Result            |
|---------------------------|---------------------|-------------------|
| ANCOVA (primary)          | adj. diff = $-3.38$ | $p = .626$        |
| Permutation test          | group coefficient   | $p = .606$        |
| Bootstrap 95% CI          | adj. diff           | $[-16.65, 12.14]$ |
| Rank-based ANCOVA         | group term          | $p = .705$        |
| Robust (Huber) regression | group term          | $p = .524$        |

*Note.* All methods indicate a small, non-significant baseline-adjusted difference whose confidence interval includes zero, confirming that the primary null result is not an artefact of a single modelling assumption.

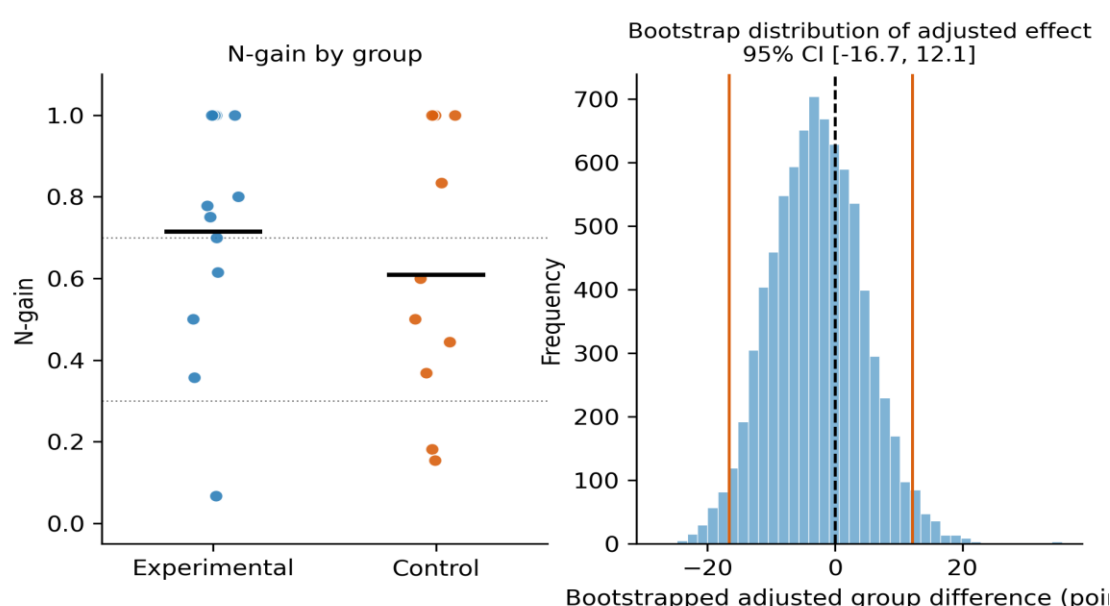


**Figure 5.** Left: normalised gain (N-gain) by group; points are individual students and the horizontal bar is the group mean, with dotted reference lines at the conventional 0.30 and 0.70 thresholds. Right: bootstrap distribution (8,000 resamples) of the baseline-adjusted between-group difference in post-test scores; the dashed line marks zero and the solid lines mark the 95% confidence limits, which straddle zero.

### Principal Finding

The study asked whether an interactive flipbook-based e-module leads to higher achievement on the properties of light than regular instruction once initial achievement is controlled. In this sample it did not. The experimental class had the higher raw post-test mean, but that advantage was already present before instruction began. After adjustment, the estimated difference was close to zero and slightly favoured the control class, and four sensitivity analyses agreed. The appropriate conclusion is that the e-module was not superior to regular instruction under the conditions studied. The result does not show that the e-module is ineffective, because both classes learned a great deal and the estimate is imprecise.

The distinction matters because both classes improved substantially, and a reader who looked only at the experimental class would conclude that the e-module "worked". A single-group pre-test–

post-test design cannot support that conclusion, since maturation, familiarity with the test, and the instruction that accompanies any material all raise post-test scores (Shadish et al., 2002). The control class experienced the same influences and gained slightly more. Only the baseline-adjusted comparison between classes isolates the contribution of the e-module, and that comparison showed no advantage.

### Comparison With Previous Research

At first sight the result conflicts with the largely positive literature on flipbook and interactive e-modules in Indonesian schools. Development studies have reported high validity and practicality together with learning gains for flipbook modules in science, vocational, and thematic subjects (Winatha, 2018; Widiana & Rosy, 2021; Munzil et al., 2022; Rohmatin et al., 2022; Sabil et al., 2023), and studies of interactive e-modules in elementary science have reported favourable student responses and better learning outcomes (Lase & Sudarma, 2023; Alyusfitri et al., 2024; Saralee et al., 2024). Most of these studies, however, judged effectiveness from expert validation or from single-group pre-test-post-test gains (Mahmudah et al., 2022; Saralee et al., 2024). In the present data, gains of more than 30 points appeared just as clearly in the class that did not use the e-module. The finding is therefore less a contradiction of earlier work than a reminder that within-group gains and validation percentages are weak evidence of superiority.

Studies that included a control class reported clear advantages for e-modules. Arina et al. (2025) compared 20 sixth-grade students who used an IPAS e-module with 20 students taught conventionally and found significantly higher post-test scores, with mean normalised gains of 0.68 and 0.29. Maulidya and Zulherman (2024) reported normalised gains of 63% and 29% for a flipbook e-module and a control class in fourth-grade IPAS, with 30 students per group. Two features distinguish those studies from the present one. Their conclusions rest on post-test comparisons and normalised gain rather than on a baseline-adjusted estimate with a confidence interval, and, as noted above, normalised gain can favour the group that starts higher (Nissen et al., 2018). In addition, the size of the reported differences suggests that the comparison instruction in those studies may have produced much smaller gains than the regular instruction in the present control class, which gained 35 points. In our own data the experimental class also had the higher normalised gain (0.71 vs. 0.61), yet this advantage disappeared once the baseline difference was modelled.

The present result is more consistent with controlled and meta-analytic research. Hillmayr et al. (2020) found that digital tools in mathematics and science generally help, but that their effects vary with how the tools are used and are smaller when the tools replace rather than supplement other instruction. Noetel et al. (2022) showed that the benefits of multimedia depend on specific design principles rather than on the digital medium as such. Students' positive perceptions of visualization technologies (Abdinejad et al., 2021) and the rapid growth of these technologies in education (Al-Ansi et al., 2023) do not by themselves establish learning gains over good conventional teaching. When the control group receives competent instruction on the same content, the point of comparison is not the absence of teaching but another form of teaching, so the added value of a particular medium is likely to be small, as is typical of effects measured against business-as-usual instruction (Kraft, 2020).

### Theoretical Interpretation

From the perspective of the cognitive theory of multimedia learning, a flipbook e-module can support learning about light by pairing verbal explanations with pictures of processes that are normally invisible, helping learners select, organise, and integrate information (Mayer & Moreno, 1998; Mayer, 2024). The theory, however, predicts an advantage of well-designed multimedia over poorly designed presentations, not an advantage of any digital material over any alternative (Noetel et al., 2022). If the control lessons also combined words and pictures through teacher explanation, textbook diagrams, or classroom demonstrations, both classes may have relied on similar processing,

leaving little room for a difference. Cognitive load theory leads to the same expectation. A page-turning digital format adds value only when it reduces extraneous load or manages intrinsic load better than the alternative (Sweller, 1988; Sweller et al., 2019), and features such as navigation, audio, and external video can add load as well as remove it. A conceptual change perspective offers a third explanation. Ideas about light are deeply rooted and change only when students are led to confront them (Fetherstonhaugh & Treagust, 1992; Bendall et al., 1993; Métioui, 2023). The module presented correct explanations with everyday pictures rather than activities that explicitly elicited and challenged students' existing ideas, so the delivery format may have mattered less than whether students were prompted to question what they already believed.

### **Alternative Explanations**

Several factors other than a true absence of effect could explain the result. The first is residual baseline imbalance. ANCOVA adjusts for the measured pre-test, but in non-randomised designs it cannot remove differences in unmeasured characteristics, and a pre-test measured with error adjusts only partly for the true difference (Shadish et al., 2002). The second is a ceiling effect. Four of 12 experimental students and 3 of 10 control students reached the maximum post-test score, which compresses variance at the top of the scale and can bias estimates and reduce power (Šimkovic & Träuble, 2019); because the experimental class started higher, the ceiling probably constrained it more. The third is confounding by teacher and class: with one class per condition, the teacher, classroom climate, and timetable cannot be separated from the intervention. The fourth is low statistical power. Given the standard error of the adjusted difference (6.82 points), the design had 80% power only for adjusted differences of about 20 points, roughly 0.9 of the pooled post-test standard deviation, so a true benefit of 5 to 10 points, which would be educationally meaningful, could easily go undetected. For the same reason the result cannot be read as evidence of equivalence, which would require the confidence interval to lie within narrow bounds of practical equivalence (Lakens, 2017). The fifth is timing: the post-test was given immediately after instruction, so differences in retention could not be observed. The fair statement is therefore that the study found no evidence that either condition was better, not that the e-module had no effect.

### **Practical Implications**

For teachers, the results suggest that an interactive flipbook e-module on the properties of light is unlikely to harm learning, since both classes learned a great deal, but that its digital and interactive format does not guarantee better test results than competent regular teaching. In practice, its effect could be negligible, a small gain, or a small loss. Decisions about adopting such a module should therefore also weigh access to devices and the internet, student engagement, opportunities for self-paced review, and teacher preparation time, none of which this study measured. Teachers may use the module where it fits, for example for revision or for students who benefit from revisiting visual explanations, without expecting it to outperform strong conventional instruction. For developers, the results argue for moving beyond feasibility scores and single-group gains toward controlled, baseline-adjusted evaluations, and for designing features that elicit and challenge common misconceptions about light instead of relying on the visual appeal of page-turning.

### **Limitations and Future Research**

Several limitations bound these conclusions. The sample was small ( $N = 22$ ) and unequal (12 vs. 10), which produced wide confidence intervals. Students were not randomised, and the baseline difference that is common between intact classes was indeed present; covariate adjustment is only a partial remedy. With one class per condition, teacher and classroom effects are confounded with the intervention. The post-test ceiling reduced variance and may have hidden differences among high achievers. No formal validity or reliability evidence was available for the achievement test, the control lessons were not documented in detail, the test measured overall achievement rather than

conceptual change or specific misconceptions, and the post-test was immediate, so retention was not assessed. Finally, the study took place in one school and on one topic, which limits generalisation.

Future studies can address these limitations in specific ways. First, sample size should be planned with a power analysis: with the pre-test as covariate and a pre-post correlation of .60 to .70, detecting a medium effect ( $d = 0.50$ ) with 80% power at  $\alpha = .05$  requires roughly 33 to 41 students per group (Cohen, 1988), and more when classes rather than students are the unit of assignment. Second, several classes per condition, ideally assigned at random, would separate the effect of the module from teacher and class effects. Third, ceiling effects can be reduced by adding more demanding items, such as items that require explanation or ray-diagram reasoning, and by checking item difficulty in a try-out. Fourth, two-tier or other diagnostic items on light would show whether a module changes misconceptions and not only total scores. Fifth, a delayed post-test would show whether any difference persists. Finally, the control condition should be documented in as much detail as the intervention, so that readers know what the e-module is being compared with.

## CONCLUSIONS

The interactive flipbook-based e-module did not improve fifth-grade students' achievement on the properties of light more than regular instruction did. Both classes improved substantially from pre-test to post-test, but after controlling for the pre-test the difference between them was small, not significant, and slightly in favour of the control class (adjusted difference =  $-3.38$  points, 95% CI  $[-17.65, 10.90]$ ). Permutation, bootstrap, rank-based, and robust analyses gave the same result. The large gain of the experimental class is therefore not evidence that the e-module was effective, because the control class gained as much without it. The small, non-randomised sample, the baseline imbalance, and the post-test ceiling mean that a moderate effect in either direction cannot be ruled out, so the finding is best described as no demonstrated superiority rather than proven ineffectiveness. Schools can use the module as one resource for learning about light, but should not adopt it in the expectation that it will raise test scores beyond those achieved by good regular teaching. For researchers, claims about the effectiveness of e-modules should rest on baseline-adjusted comparisons between groups reported with effect sizes and confidence intervals. A definitive answer will require larger multi-class samples, outcome measures that are more demanding and diagnostic of misconceptions, and a delayed post-test.

## ACKNOWLEDGMENTS

The authors gratefully acknowledge Universitas Nusantara PGRI Kediri for the support and facilities that made this research possible. The authors also thank the principal, teachers, and fifth-grade students of Grogol 3 Public Elementary School for their cooperation and participation throughout the implementation of the study.

## REFERENCES

- Abdinejad, M., Talaie, B., Qorbani, H. S., & Dalili, S. (2021). Student perceptions using augmented reality and 3D visualization technologies in chemistry education. *Journal of Science Education and Technology*, 30(1), 87–96. <https://doi.org/10.1007/s10956-020-09880-2>
- Al-Ansi, A. M., Jaboob, M., Garad, A., & Al-Ansi, A. (2023). Analyzing augmented reality (AR) and virtual reality (VR) recent development in education. *Social Sciences & Humanities Open*, 8(1), Article 100532. <https://doi.org/10.1016/j.ssaho.2023.100532>
- Alyusfitri, R., Gistituati, N., Yerizon, Fauzan, A., & Yarman. (2024). The effectiveness and relationship of student responses toward learning outcomes using interactive multimedia-

- based e-modules in elementary schools. *International Electronic Journal of Elementary Education*, 16(5), 573–584. <https://iejee.com/index.php/IEJEE/article/view/2203>
- American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). American Psychological Association. <https://doi.org/10.1037/0000165-000>
- Arina, N., Mustaji, Kholidya, C. F., Rulviana, V., Rapii, M., & Wardana, N. A. D. (2025). The effectiveness of e-module use on the science learning outcomes of sixth-grade primary school students. *Jurnal Penelitian Pendidikan IPA*, 11(12), 1381–1389. <https://doi.org/10.29303/jppipa.v11i12.13593>
- Bendall, S., Goldberg, F., & Galili, I. (1993). Prospective elementary teachers' prior knowledge about light. *Journal of Research in Science Teaching*, 30(9), 1169–1187. <https://doi.org/10.1002/tea.3660300912>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Fetherstonhaugh, T., & Treagust, D. F. (1992). Students' understanding of light and its properties: Teaching to engender conceptual change. *Science Education*, 76(6), 653–672. <https://doi.org/10.1002/sce.3730760606>
- 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>
- Hillmayr, D., Ziernwald, L., Reinhold, F., Hofer, S. I., & Reiss, K. M. (2020). The potential of digital tools to enhance mathematics and science learning in secondary schools: A context-specific meta-analysis. *Computers & Education*, 153, Article 103897. <https://doi.org/10.1016/j.compedu.2020.103897>
- Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, 49(4), 241–253. <https://doi.org/10.3102/0013189X20912798>
- Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4), 355–362. <https://doi.org/10.1177/1948550617697177>
- Lase, K. S., & Sudarma, I. K. (2023). Interactive e-module based on scientific approach in science content learning for sixth-grade students. *MIMBAR PGSD Undiksha*, 11(2), 175–183. <https://doi.org/10.23887/jjpgsd.v11i2.64508>
- Mahmudah, S., Kirana, T., & Rahayu, Y. S. (2022). Profile of students' critical thinking ability: Implementation of e-modul based on problem-based learning. *IJORER: International Journal of Recent Educational Research*, 3(4), 478–488. <https://doi.org/10.46245/ijorer.v3i4.231>
- Maulidya, S. A., & Zulherman. (2024). 1stFlip flipbook creator-assisted e-module to advance learning quality in elementary schools. *International Journal of Elementary Education*, 8(2), 249–257. <https://doi.org/10.23887/ijee.v8i2.78076>
- Mayer, R. E. (2024). The past, present, and future of the cognitive theory of multimedia learning. *Educational Psychology Review*, 36, Article 8. <https://doi.org/10.1007/s10648-023-09842-1>
- Mayer, R. E., & Moreno, R. (1998). A split-attention effect in multimedia learning: Evidence for dual processing systems in working memory. *Journal of Educational Psychology*, 90(2), 312–320. <https://doi.org/10.1037/0022-0663.90.2.312>
- Métioui, A. (2023). Primary school preservice teachers' alternative conceptions about light interaction with matter (reflection, refraction, and absorption) and shadow size changes on Earth and Sun. *Education Sciences*, 13(5), Article 462. <https://doi.org/10.3390/educsci13050462>
- Munzil, M., Affriyenni, Y., Mualifah, S., Fardhani, I., Fitriyah, I. J., & Muntholib, M. (2022). Development of problem based learning based e-modules in the form of flipbooks on environmentally friendly technology materials. *Jurnal Pendidikan Sains Indonesia*, 10(1), 37–46. <https://doi.org/10.24815/jpsi.v10i1.21807>

- Nissen, J. M., Talbot, R. M., Nasim Thompson, A., & Van Dusen, B. (2018). Comparison of normalized gain and Cohen's d for analyzing gains on concept inventories. *Physical Review Physics Education Research*, 14(1), Article 010115. <https://doi.org/10.1103/PhysRevPhysEducRes.14.010115>
- 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/10.3102/00346543211052329>
- Rohmatin, I. A., Rachmayani, A., & Jumadi, J. (2022). Development of e-module based on flipbook learning model problem based learning (PBL) to improve critical thinking ability. *Berkala Ilmiah Pendidikan Fisika*, 10(3), 342–352. <https://doi.org/10.20527/bipf.v10i3.13655>
- Sabil, H., Asrial, A., Syahrial, S., Kiska, N. D., Saputri, J., Damayanti, L., Luthfiah, Q., & Silvia, N. (2023). Problem-based learning flipbook e-module in improving students' critical thinking skills in "Always Save Energy" thematic learning. *International Journal of Elementary Education*, 7(1). <https://doi.org/10.23887/ijee.v7i1.58306>
- Saralee, A., Fitria, Y., Sukma, E., & Hidayati, A. (2024). The development of e-module IPAS based on problem based learning assisted with articulate applications to improve students' critical thinking. *Jurnal Penelitian Pendidikan IPA*, 10(4), 2032–2040. <https://doi.org/10.29303/jppipa.v10i4.7014>
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.
- Šimkovic, M., & Träuble, B. (2019). Robustness of statistical methods when measure is affected by ceiling and/or floor effect. *PLOS ONE*, 14(8), Article e0220889. <https://doi.org/10.1371/journal.pone.0220889>
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285. [https://doi.org/10.1207/s15516709cog1202\\_4](https://doi.org/10.1207/s15516709cog1202_4)
- 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>
- Vickers, A. J., & Altman, D. G. (2001). Analysing controlled trials with baseline and follow up measurements. *BMJ*, 323(7321), 1123–1124. <https://doi.org/10.1136/bmj.323.7321.1123>
- Widiana, F. H., & Rosy, B. (2021). Pengembangan e-modul berbasis flipbook maker pada mata pelajaran teknologi perkantoran. *Edukatif: Jurnal Ilmu Pendidikan*, 3(6), 3728–3739. <https://doi.org/10.31004/edukatif.v3i6.1265>
- Winatha, K. R. (2018). Pengembangan e-modul interaktif berbasis proyek mata pelajaran simulasi digital. *Jurnal Pendidikan Teknologi dan Kejuruan*, 15(2), 188–199. <https://doi.org/10.23887/jptk-undiksha.v15i2.14021>