EDUKASI
Published:
Faculty of Teacher Training and Education
Universitas Khairun

e-ISSN 2715-8551

DOI: http://dx.doi.org/10.33387/j.edu.v24i2.13088
https://ejournal.unkhair.ac.id/index.php/edu
(Accredited Sinta Rank 3)

Index by:

Augmented reality versus two-dimensional media in problem-based learning: Effects on elementary students' conceptual understanding of Earth and space

Rahayu Ramadanaryanthi[1], Said Hasan[2], Zulkifli Ahmad[3]

[1,2,3] Master's Biology Education Study Program, FKIP, Universitas Khairun, Ternate, Indonesia
Email: rahayuramadanaryanthi96@guru.sd.belajar.id; saidhasan@unkhair.ac.id; zul_bio@unkhair.ac.id

ARTICLE INFO

Keywords:
Augmented reality;
Problem-based learning;
Conceptual understanding;
Earth and space;
Elementary school;

Article history:
Received 2026-08-08
Revised 2026-09-10
Accepted 2026-10-30

ABSTRACT

Earth and space concepts are difficult for elementary students because the phenomena are abstract, spatial, and rarely observable. This study examined whether Assemblr Edu-based augmented reality (AR) embedded in problem-based learning (PBL) improves sixth graders' conceptual understanding more than identical PBL using two-dimensional (2D) media. A quasi-experimental pretest–posttest nonequivalent control group design involved two intact classes at a public elementary school in Ternate, Indonesia (n = 25 each). A 10-item essay test measured six cognitive processes of understanding. Because some distributions were non-normal, normalized gain (N-Gain) was compared with the Mann–Whitney U test, supported by Wilcoxon tests, ANCOVA, and Quade's rank ANCOVA. The experimental class improved from 62.00 to 91.20 (N-Gain = 0.78, high) and the control class from 54.80 to 74.80 (N-Gain = 0.43, moderate). N-Gain was significantly higher with AR ($U = 4.5$, $p < .001$, $r = .85$, Cliff's $\delta = .99$), and the pretest-adjusted posttest advantage persisted ($F(1, 47) = 131.14$, $p < .001$, $\eta p^2 = .74$). AR was superior on five of six indicators, notably inferring and comparing, but not interpreting. AR thus adds value over 2D media within PBL when 3D models serve as evidence for spatial reasoning; the unusually large effect requires confirmation.

This is an open access article under the CC BY-NC-SA license.

Corresponding Author:
Rahayu Ramadanaryanthi
Program Studi Magister Pendidikan Biologi, FKIP, Universitas Khairun;
rahayuramadanaryanthi96@guru.sd.belajar.id

EDUKASI
Published:
Faculty of Teacher Training and Education
Universitas Khairun

e-ISSN 2715-8551

DOI: http://dx.doi.org/10.33387/j.edu.v24i2.13088
https://ejournal.unkhair.ac.id/index.php/edu
(Accredited Sinta Rank 3)

Index by:

INTRODUCTION

Science and social studies (IPAS) instruction in Indonesian elementary schools aims to develop scientific literacy and critical reasoning from an early age. Conceptual understanding underpins both goals, because students cannot reason scientifically about phenomena they do not yet understand. In practice, however, elementary students' understanding of science concepts remains problematic, particularly in astronomy. Putri et al. (2021) found that 88 sixth-grade students in Pekanbaru held moderate levels of misconceptions about the solar system, with an average of 47.58%. Similar findings were reported by Fitriyani et al. (2025) using a four-tier test and by Amelia et al. (2025) among sixth graders, indicating that astronomy misconceptions are persistent and cannot be resolved through verbal explanation alone.

The topic "Exploring Earth and Space" covers Earth's rotation and revolution, day and night, time differences between regions, seasons, lunar phases, and eclipses. All of these concepts require spatial reasoning, that is, the ability to imagine the positions and motions of the Sun, Earth, and Moon from a viewpoint outside Earth and to relate them to what is observed from Earth's surface. These phenomena occur at spatial and temporal scales that cannot be observed directly in the classroom. When instruction relies only on text and static two-dimensional (2D) pictures, students must reconstruct three-dimensional (3D) models in their minds. This process taxes working memory and increases the likelihood of inaccurate mental models (Sweller et al., 2019).

Augmented reality (AR) offers a way to bridge this representational gap. AR combines real and virtual objects in a real environment, registers them with each other in three dimensions, and allows interaction in real time (Azuma, 1997). This allows students to rotate, enlarge, and observe a 3D model from multiple angles. In Ternate, Assemblr Edu-based AR has already been used to improve senior high school students' learning outcomes (Ahmad et al., 2022). Systematic reviews show that AR improves learning achievement, motivation, and understanding of abstract concepts, although its implementation also faces technical and pedagogical challenges (Akçayır & Akçayır, 2017; Ibáñez & Delgado-Kloos, 2018). A meta-analysis of 64 quantitative studies (N = 4,705) reported an effect of AR on learning gains of d = 0.68 (Garzón & Acevedo, 2019), and a recent review of AR in science learning found that most studies used quasi-experimental designs and reported generally positive results (Zufahmi, 2025). Theoretically, the benefits of AR can be explained by the Cognitive Theory of Multimedia Learning, which emphasizes integrated visual and verbal representations (Mayer, 2024), and by Cognitive Load Theory, which highlights the limits of working memory when processing information with high element interactivity (Sweller et al., 2019).

Empirical evidence in astronomy is also promising. Tran (2024) reported that 438 sixth-grade lower-secondary students in Vietnam who learned with mobile AR outperformed a control group in celestial motion, spatial orientation, and conceptual understanding. Ferrari

EDUKASI
Published:
Faculty of Teacher Training and Education
Universitas Khairun
e-ISSN 2715-8551
DOI: http://dx.doi.org/10.33387/j.edu.v24i2.13088
https://ejournal.unkhair.ac.id/index.php/edu
(Accredited Sinta Rank 3)
Index by:

et al. (2024) found large effects of AR on astronomy learning in secondary school, and Limpinan et al. (2026) showed that second graders who learned about the solar system with AR or virtual reality outperformed peers taught conventionally.

Visual technology, however, does not automatically produce understanding. Wen et al. (2023) found that AR did not produce a significant difference in the academic performance of primary students in Singapore, although combining AR with an inquiry framework improved critical thinking and knowledge-creation efficacy. Faria (2024) even found that adding AR to conventional teaching did not improve learning and may have caused cognitive overload, concluding that the teaching strategy matters as much as the technology itself. These findings indicate that AR must be embedded in an instructional design that gives students a reason to observe, compare, and draw conclusions from virtual objects.

Problem-based learning (PBL) is one framework suited to this purpose. A meta-analysis of 55 experimental studies showed that PBL has a large effect on science achievement compared with traditional instruction, and that this effect is relatively consistent across educational levels (Uluçınar, 2023). Because the PBL syntax begins with an authentic problem and proceeds through group investigation, it gives AR a clear function: a source of evidence that students explore to solve the problem. Arici and Yilmaz (2023) showed that AR-supported PBL outperformed both PBL alone and conventional instruction in academic achievement, reflective thinking, and decision-making among seventh graders. In Indonesia, Gibran and Setyasto (2025) reported that AR-assisted PBL produced a moderate N-Gain (0.43) among fifth graders, whereas picture-assisted PBL produced almost no gain.

These studies leave several gaps. First, AR research on astronomy has been conducted mostly at the secondary level, and studies with elementary students remain limited. Second, AR-PBL studies generally measure overall achievement and have not mapped conceptual understanding according to the cognitive processes of understanding in Anderson and Krathwohl's (2001) taxonomy: interpreting, exemplifying, classifying, inferring, comparing, and explaining. Third, many studies compare AR with conventional instruction, making it difficult to separate the contribution of the medium from that of the instructional model. This study addresses all three gaps. It involves sixth-grade elementary students (first gap), maps gains across the six cognitive processes of understanding (second gap), and applies the same PBL syntax in both classes while varying only the medium, namely 3D AR versus 2D visual media (third gap). The study was also conducted in Ternate, North Maluku, adding evidence from schools in eastern Indonesia.

Accordingly, this study aimed to: (1) examine differences in gains in conceptual understanding of Earth and space between students taught with Assemblr Edu-based AR in PBL and students taught with 2D visual media in PBL; and (2) describe the gains in each cognitive process of understanding. The research hypothesis was directional: N-Gain in conceptual understanding was expected to be higher in the AR-supported PBL class than in

EDUKASI
Published:
Faculty of Teacher Training and Education
Universitas Khairun
e-ISSN 2715-8551
DOI: http://dx.doi.org/10.33387/j.edu.v24i2.13088
https://ejournal.unkhair.ac.id/index.php/edu
(Accredited Sinta Rank 3)
Index by:

the 2D-supported PBL class. To keep the test conservative, it was evaluated with two-tailed tests at $\alpha = .05$ (H_0: the N-Gain distributions of the two classes do not differ; H_1: they differ). The hypothesis was considered supported only if the difference was significant and favoured the AR class.

METHOD

Design and participants

This quantitative study used a quasi-experimental pretest–posttest nonequivalent control group design. This design was chosen because students were taught in intact classes, which made individual random assignment impossible. The study was conducted at SD Negeri 2 Kota Ternate, North Maluku, Indonesia, from February to May 2026 (second semester of the 2025/2026 academic year). This period covered the whole research-and-development cycle: preparation of the instruments, development and expert validation of the AR media, a limited tryout of the media with one IPAS teacher and 10 sixth-grade students, classroom implementation, data collection, and analysis. The classroom phase itself consisted of the pretest, two 30-minute intervention sessions, and the posttest . The researcher, who is a teacher at the school, taught both classes. This kept the instructor constant across conditions, but it also means that the instructor knew the study hypothesis, a point addressed in the limitations. Participants were 50 sixth-grade students from two intact classes: class VI-A ($n = 25$) served as the experimental class and class VI-B ($n = 25$) as the control class. Both classes used the same teaching module, time allocation, and test.

Ethical considerations. The study did not receive formal approval from a research ethics committee. It was conducted as part of regular IPAS lessons with the permission of the school principal. Parents or guardians were informed of the study orally and through a message in the class parents' WhatsApp group, and they gave consent through that group; no signed consent forms were collected. Students were told orally that their test answers would be used for research and would not affect their report-card grades, and they gave oral assent. Students were identified only by code numbers, the data were stored on a password-protected device accessible only to the research team, and only group-level results are reported.

Media and intervention procedure

The AR media were developed with Assemblr Edu following the ADDIE model as part of a broader research-and-development project. The media included 3D models of the Earth, Sun, and Moon, visualizations of Earth's rotation and revolution, and the parts of the Sun. Before use, the media were validated by two media experts (mean 96.92%) and two content experts (mean 94.80%), both in the very valid category. Practicality testing yielded scores of 96.07% from student responses and 97.65% from teacher responses (very practical).

EDUKASI
Published:
Faculty of Teacher Training and Education
Universitas Khairun
e-ISSN 2715-8551
DOI: http://dx.doi.org/10.33387/j.edu.v24i2.13088
https://ejournal.unkhair.ac.id/index.php/edu
(Accredited Sinta Rank 3)
Index by:

Both classes followed the same five PBL phases. The only difference was the medium students used during the investigation phase: the experimental class used AR on Android devices, whereas the control class used pictures, posters, and 2D worksheets with equivalent content. In both classes, students worked in small groups formed in the same way. Each experimental group shared one Android smartphone running the Assemblr Edu application (the version available during data collection in 2026). Students scanned AR markers printed on their worksheets or displayed on the classroom digital board. Each control group received one identical set of 2D materials, so every group had a single shared resource for the investigation. The intervention lasted two sessions of 30 minutes each, excluding the pretest and posttest. Table 1 details the activities in each phase.

Table 1. Learning activities in each PBL phase in the experimental and control classes

PBL phase	Experimental class (PBL + AR)	Control class (PBL + 2D media)
Orienting students to the problem	The teacher presents an authentic problem about time differences and the day–night cycle.	Same as the experimental class.
Organizing students for learning	Students work in groups and receive a teaching module containing AR markers.	Students work in groups and receive a teaching module with 2D pictures.
Guiding investigation	Students scan the markers with Assemblr Edu and explore 3D models of rotation, revolution, axial tilt, and sunlight.	Students observe 2D pictures, diagrams, and posters of rotation, revolution, and axial tilt.
Developing and presenting results	Groups complete an investigation sheet and present their scientific arguments.	Same as the experimental class.
Analyzing and evaluating	The teacher facilitates reflection to correct misconceptions and draws conclusions with the students.	Same as the experimental class.

Instrument

Conceptual understanding was measured with a 10-item essay test based on the cognitive process dimension of understanding in Anderson and Krathwohl's (2001) taxonomy. Items were distributed across indicators as follows: interpreting (items 1–2), exemplifying (item 3), classifying (items 4–5), inferring (items 6–7), comparing (item 8), and explaining (items 9–10), according to the test blueprint. Each answer was scored from 0 to 4 using an analytic rubric, giving a total score of 0–40 that was converted to a 0–100 scale. Content validity was established through expert judgment of the alignment between each item and its indicator in the test blueprint. This judgment was qualitative; no quantitative content validity index, such as Aiken's V, was computed, and this is acknowledged as a limitation. No separate item tryout was conducted before the study. Item statistics computed post hoc in the study sample showed difficulty indices (mean item score divided by 4) of .52–.67 at pretest and .76–.92 at posttest, and corrected item–total correlations of .43–.83 and .42–.85, respectively. All items therefore discriminated adequately, but the posttest items were easy for many students, which is consistent with the ceiling effect noted in the limitations. Internal consistency in the study

EDUKASI
Published:
Faculty of Teacher Training and Education
Universitas Khairun
e-ISSN 2715-8551
DOI: http://dx.doi.org/10.33387/j.edu.v24i2.13088
https://ejournal.unkhair.ac.id/index.php/edu
(Accredited Sinta Rank 3)
Index by:

sample (N = 50) was high, with Cronbach's α = .90 at pretest and .89 at posttest. All answers were scored by the researcher using the same analytic rubric for both classes. The scorer knew which class each student belonged to, and inter-rater reliability was not estimated; this is acknowledged as a limitation.

Data analysis

Each student's gain in conceptual understanding was calculated as the normalized gain (N-Gain) using the formula of Hake (1998). Hake defined the gain at the class level from class mean scores (gain of averages). In this study the gain was computed for each student and then averaged (average of individual gains), which allows individual gains to be compared between classes with nonparametric tests. The two approaches can yield slightly different values; based on the class means in Table 2, the class-level gains were 0.77 for the experimental class and 0.44 for the control class, compared with average individual gains of 0.78 and 0.43:

$$g = (S_{post} - S_{pre}) / (S_{max} - S_{pre})$$

where S_{max} = 100. Values of $g \geq 0.70$ were classified as high, $0.30 \leq g < 0.70$ as moderate, and $g < 0.30$ as low (Hake, 1998). N-Gain is undefined when $S_{pre} = S_{max}$, so such cases were excluded from the gain analysis of the affected score. No student reached the maximum total pretest score (highest = 80). At indicator level this occurred once, for comparing in the experimental class, so that comparison used n = 24.

The analysis proceeded in four steps. First, data were described using means, standard deviations, medians, interquartile ranges (IQR), and minimum and maximum scores. Second, normality was tested with the Shapiro–Wilk test and homogeneity of variance with the median-based Levene test. Third, baseline equivalence was tested on pretest scores, and the hypothesis was tested by comparing N-Gain between classes using the Mann–Whitney U test, because some data were not normally distributed. Within-class gains were tested with the Wilcoxon signed-rank test. Because N-Gain only partly corrects for baseline differences and can be biased when pretest scores are high, ANCOVA with pretest scores as the covariate and Quade's rank analysis of covariance (Quade, 1967) were treated as the main supporting tests of the hypothesis. Fourth, N-Gain for each indicator was compared between classes using the Mann–Whitney U test with Holm's correction for multiple comparisons (Holm, 1979).

Effect sizes for nonparametric tests were calculated as $r = |Z| / \sqrt{N}$ (Fritz et al., 2012), where N is the total number of students for the Mann–Whitney test. For the Wilcoxon test, N was set to the number of pairs (25), because each student contributes one difference score. Because some studies instead use the total number of observations (2 × 25 = 50; Fritz et al., 2012), both values are reported to allow comparison. Values of r of .10, .30, and .50 were interpreted as small, medium, and large effects (Cohen, 1988). For between-class comparisons, Cliff's δ (Cliff, 1993) and the Hodges–Lehmann estimate of the median difference with a bootstrap 95% confidence interval (10,000 resamples, percentile method, random seed = 2026) are also

EDUKASI
Published:
Faculty of Teacher Training and Education
Universitas Khairun
e-ISSN 2715-8551
DOI: http://dx.doi.org/10.33387/j.edu.v24i2.13088
https://ejournal.unkhair.ac.id/index.php/edu
(Accredited Sinta Rank 3)
Index by:

reported; for ANCOVA, partial eta squared (η_p^2) is reported. Z values were corrected for ties. All tests used $\alpha = .05$. For this revision, all analyses were re-run from the anonymized item-level data in Python 3.12.3 with SciPy 1.17.1, NumPy 2.4.4, and pandas 3.0.2. The ANCOVA was estimated by ordinary least squares, Quade's rank ANCOVA was computed as a one-way ANOVA on residuals from the regression of posttest ranks on pretest ranks, and Cliff's δ and the bootstrap interval were computed directly. The re-analysis reproduced all previously reported values.

RESULTS AND DISCUSSION

Results

Descriptive statistics. Table 2 presents descriptive statistics for pretest scores, posttest scores, and N-Gain. The experimental class improved from a mean of 62.00 to 91.20, with an N-Gain of 0.78 (high), whereas the control class improved from 54.80 to 74.80, with an N-Gain of 0.43 (moderate). In the experimental class, 19 students (76%) achieved a high N-Gain and 6 (24%) a moderate N-Gain. In the control class, no student achieved a high N-Gain; 21 students (84%) were in the moderate category and 4 (16%) in the low category. Four students in the experimental class obtained the maximum posttest score (100), whereas the highest score in the control class was 80.

Table 2. Descriptive statistics for conceptual understanding scores and N-Gain

Class	Variable	*M* (*SD*)	*Mdn* (IQR)	Min–Max
Experimental (n = 25)	Pretest	62.00 (8.16)	60.00 (10.00)	50–80
	Posttest	91.20 (4.40)	90.00 (0.00)	80–100
	N-Gain	0.78 (0.11)	0.75 (0.05)	0.60–1.00
Control (n = 25)	Pretest	54.80 (9.18)	60.00 (10.00)	40–70
	Posttest	74.80 (5.10)	70.00 (10.00)	70–80
	N-Gain	0.43 (0.12)	0.40 (0.17)	0.25–0.67

Note. *M* = mean; *SD* = standard deviation; *Mdn* = median; IQR = interquartile range. Maximum possible score = 100. N-Gain is the mean of individual N-Gain scores.

Assumption tests. The Shapiro–Wilk test (Table 3) showed that pretest and posttest scores in both classes, as well as N-Gain in the experimental class, were not normally distributed ($p < .05$); only N-Gain in the control class was normally distributed ($p = .059$). Variances were homogeneous for pretest scores and N-Gain but not for posttest scores ($p = .037$). Therefore, the main hypothesis was tested with nonparametric statistics.

Table 3. Shapiro–Wilk normality tests and Levene's tests of homogeneity

Data	Experimental *W* (*p*)	Control *W* (*p*)	Levene *F* (*p*)
Pretest	0.866 (.004)	0.882 (.008)	0.747 (.392)
Posttest	0.597 (< .001)	0.639 (< .001)	4.594 (.037)
N-Gain	0.808 (< .001)	0.923 (.059)	1.105 (.298)

EDUKASI
Published:
Faculty of Teacher Training and Education
Universitas Khairun
e-ISSN 2715-8551
DOI: http://dx.doi.org/10.33387/j.edu.v24i2.13088
https://ejournal.unkhair.ac.id/index.php/edu
(Accredited Sinta Rank 3)
Index by:

Note. Levene's test used the median as the center (Brown–Forsythe variant); $df = 1, 48$.

Baseline equivalence. The Mann–Whitney U test showed that pretest scores were significantly higher in the experimental class than in the control class ($U = 185.0$, $Z = -2.611$, $p = .009$, $r = .37$). Because the classes were not equivalent at baseline, N-Gain was used as the primary outcome. N-Gain, however, only partly corrects for initial differences and may be biased when pretest scores are high (ceiling effects), so the ANCOVA and Quade's rank ANCOVA reported below serve as the main supporting tests.

Hypothesis testing. Table 4 summarizes the within- and between-class tests. Both classes improved significantly from pretest to posttest with large effects (experimental: $Z = -4.497$, $p < .001$, $r = .90$ [$r = .64$ with $N = 50$ observations]; control: $Z = -4.446$, $p < .001$, $r = .89$ [$r = .63$]); every student in both classes scored higher on the posttest than on the pretest. The between-class comparison showed that N-Gain was significantly higher in the experimental class than in the control class ($U = 4.5$, $Z = -6.041$, $p < .001$, $r = .85$). Cliff's δ of .99 indicates that almost every student in the experimental class had a higher N-Gain than any student in the control class, with an estimated median difference of 0.35 (95% CI [0.25, 0.42]). The research hypothesis was therefore supported.

Table 4. Wilcoxon signed-rank tests (within classes) and Mann–Whitney U tests (between classes)

Comparison	Test statistic	*Z*	*p*	*r*	Effect size
Pretest–posttest, experimental	$T = 0.0$	−4.497	< .001	.90 (.64)	Large
Pretest–posttest, control	$T = 0.0$	−4.446	< .001	.89 (.63)	Large
Pretest, experimental vs. control	$U = 185.0$	−2.611	.009	.37	Medium
N-Gain, experimental vs. control	$U = 4.5$	−6.041	< .001	.85	Large

Note. $r = |Z|/\sqrt{N}$; $N = 25$ (number of pairs) for the Wilcoxon tests and, in parentheses, $N = 50$ (total observations; Fritz et al., 2012); $N = 50$ for the Mann–Whitney tests. Z values are corrected for ties.

Robustness analyses. After controlling for pretest scores, the adjusted posttest mean of the experimental class (90.14) remained higher than that of the control class (75.86), a difference of 14.29 points (95% CI [11.78, 16.80]; $F(1, 47) = 131.14$, $p < .001$, $\eta_p^2 = .74$). The assumption of homogeneous regression slopes was met (class × pretest interaction: $p = .569$), and the residuals were normally distributed (Shapiro–Wilk $p = .074$). Because posttest variances were not homogeneous, the result was confirmed with Quade's rank ANCOVA, which led to the same conclusion ($F(1, 48) = 104.15$, $p < .001$). Thus, the advantage of the experimental class cannot be explained by baseline differences.

Gains by indicator. Table 5 presents mean scores and N-Gain for each indicator. The experimental class achieved significantly higher N-Gain on five of the six indicators. The largest differences were for inferring ($r = .88$) and comparing ($r = .87$), followed by exemplifying, explaining, and classifying. For interpreting, the difference between classes was not significant (Holm-adjusted $p = .434$). Exemplifying and comparing were each measured by

EDUKASI
Published:
Faculty of Teacher Training and Education
Universitas Khairun
e-ISSN 2715-8551
DOI: http://dx.doi.org/10.33387/j.edu.v24i2.13088
https://ejournal.unkhair.ac.id/index.php/edu
(Accredited Sinta Rank 3)
Index by:

a single item, so their scores are less reliable than those of the other indicators and their large differences should be interpreted with caution.

Table 5. Mean scores (%) and N-Gain by indicator of conceptual understanding in both classes

Indicator (items)	Experimental: pre → post (g)	Control: pre → post (g)	U (Z)	p (Holm)	r
Interpreting (1–2)	66.0 → 88.0 (0.62)	59.5 → 79.5 (0.48)	273.0 (−0.78)	.434	.11
Exemplifying (3)	60.0 → 92.0 (0.78)	54.0 → 62.0 (0.16)	69.5 (−5.03)	< .001	.71
Classifying (4–5)	58.5 → 89.0 (0.73)	51.0 → 75.0 (0.46)	87.5 (−4.67)	< .001	.66
Inferring (6–7)	55.0 → 89.5 (0.77)	52.5 → 75.0 (0.46)	10.5 (−6.21)	< .001	.88
Comparing (8)	59.0 → 99.0 (0.98)	46.0 → 76.0 (0.55)	26.5 (−6.10)	< .001	.87
Explaining (9–10)	71.0 → 94.0 (0.80)	61.0 → 75.5 (0.28)	76.5 (−4.78)	< .001	.68

Note. Indicator scores are expressed as percentages of the maximum indicator score. g = mean individual N-Gain. For comparing, one experimental student obtained the maximum pretest score, so N-Gain could not be calculated for that student (n = 24). Z values (in parentheses) are corrected for ties; $r = |Z|/\sqrt{N}$ with N = 50 (49 for comparing). Exemplifying and comparing are single-item indicators and should be interpreted with caution.

Discussion

This study shows that AR-supported PBL produced greater gains in conceptual understanding of Earth and space than PBL supported by 2D media. The N-Gain of the experimental class (0.78) was almost twice that of the control class (0.43), and this advantage persisted after baseline differences were controlled. Because both classes followed the same PBL syntax, module, and test, the difference can be attributed mainly to the medium, although a quasi-experimental design cannot fully rule out other class-related factors. This finding is consistent with Arici and Yilmaz (2023), who found that AR-supported PBL outperformed PBL alone, and with Gibran and Setyasto (2025), who reported an N-Gain of 0.43 for AR-assisted PBL among fifth graders. The higher N-Gain in the present study may reflect the strongly spatial nature of astronomy content, which benefits more from 3D representations. The direction of the effect agrees with the meta-analysis by Garzón and Acevedo (2019) and with AR astronomy studies by Tran (2024) and Ferrari et al. (2024), but its size does not. The effects observed here (r = .85, η_p^2 = .74, Cliff's δ = .99) are far larger than the meta-analytic average of d = 0.68 and should not be read as a typical effect of AR. Several features of the study may have inflated them: the small sample, a short intervention measured immediately afterwards, narrow score distributions with ceiling effects that shrink variability and therefore enlarge standardized effects, scoring by a single rater who knew class membership, and a possible novelty effect among students using AR for the first time. Larger and longer studies are needed to estimate the size of the effect more realistically.

EDUKASI
Published:
Faculty of Teacher Training and Education
Universitas Khairun
e-ISSN 2715-8551
DOI: http://dx.doi.org/10.33387/j.edu.v24i2.13088
https://ejournal.unkhair.ac.id/index.php/edu
(Accredited Sinta Rank 3)
Index by:

The advantage of AR can be explained by the Cognitive Theory of Multimedia Learning. Understanding day and night, time differences, and seasons requires students to integrate the direction of rotation, the tilt of Earth's axis, and the direction of incoming sunlight simultaneously. With 2D media, this information is spread across several pictures, and students must combine it mentally. The 3D models in AR present all three in a single representation that can be rotated and viewed from different angles, consistent with the spatial contiguity principle (Mayer, 2024). From the perspective of Cognitive Load Theory, AR does not change the intrinsic complexity of the content, but it reduces the extraneous load that arises when students must imagine 3D motion from static pictures (Sweller et al., 2019). This explanation remains inferential, because cognitive load was not measured directly.

The indicator-level pattern supports this explanation. The largest advantages of the experimental class were in inferring and comparing, two processes that require students to reason about the positions and motions of celestial bodies, for example comparing day and night in two regions or inferring the cause of an eclipse from the alignment of the Sun, Earth, and Moon. Explaining also showed a large difference: the control class achieved an N-Gain of only 0.28 (low), compared with 0.80 in the experimental class. In contrast, interpreting did not differ significantly, possibly because interpreting simple pictures or statements is already adequately supported by 2D media. This pattern is consistent with Putri et al. (2021) and Fitriyani et al. (2025), who found that elementary students' misconceptions center on the positions and motions of celestial bodies rather than on recognizing the objects themselves.

The role of PBL should not be overlooked. The control class also improved significantly, indicating that the PBL syntax itself contributed to conceptual understanding. Wen et al. (2023) and Faria (2024) caution that AR without an appropriate teaching strategy does not necessarily improve achievement. In this study, AR objects were used in the investigation phase to answer a problem posed at the outset, so students used the 3D models as evidence to test conjectures and build group arguments. This is consistent with Uluçınar's (2023) meta-analysis showing strong effects of PBL in science learning, and it indicates that AR added value over 2D media within the same PBL framework. Because no class learned without PBL, the study cannot separate the contribution of PBL itself from that of the medium.

This study has several limitations. *First*, the two classes were not equivalent at pretest; although this was addressed through N-Gain and ANCOVA, unmeasured class characteristics may still have influenced the results. *Second*, the sample came from a single school with 25 students per class, which limits generalizability. No a priori power analysis was conducted. A sensitivity analysis showed that, with 25 students per class, $\alpha = .05$ (two-tailed), and power of .80, the design could reliably detect only large effects ($d \approx 0.81$, or about 0.83 for the Mann–Whitney test), and its power to detect the meta-analytic average of $d = 0.68$ was only about .65. The observed effect was far above this threshold, but smaller and more realistic effects could have been missed, for example at indicator level. *Third*, the intervention was short (two 30-

EDUKASI
Published:
Faculty of Teacher Training and Education
Universitas Khairun
e-ISSN 2715-8551
DOI: http://dx.doi.org/10.33387/j.edu.v24i2.13088
https://ejournal.unkhair.ac.id/index.php/edu
(Accredited Sinta Rank 3)
Index by:

minute sessions), so the durability of the effect is unknown. *Fourth*, the score distribution was very narrow (for example, most experimental students scored 90 on the posttest) and four students reached the maximum score, so the test was less sensitive in distinguishing high-performing students. *Fifth*, there was no delayed posttest, so long-term retention and possible novelty effects could not be evaluated, and cognitive load and engagement were not measured directly. *Sixth*, all answers were scored by a single rater who knew the students' class membership, and inter-rater reliability was not estimated, so scoring bias cannot be ruled out. *Seventh*, the researcher taught both classes and knew the study hypothesis. Teaching both classes kept the instructor constant, but expectancy effects, such as more intensive interaction or greater enthusiasm in the experimental class, cannot be ruled out. *Eighth*, content validity of the test rests on qualitative expert judgment without a quantitative index. *Ninth*, the study did not obtain formal ethical clearance, and parental consent was given orally and through a WhatsApp group rather than through signed forms. Future studies should involve several schools, use longer interventions and items with more varied difficulty, include a delayed posttest, use blind double scoring of a subsample with Cohen's κ or ICC, and measure cognitive load.

CONCLUSION

Assemblr Edu-based AR integrated into PBL improved sixth-grade students' conceptual understanding of Earth and space more than PBL supported by 2D visual media. The experimental class achieved an N-Gain of 0.78 (high), whereas the control class achieved 0.43 (moderate). This difference was significant with a large effect ($U = 4.5$, $p < .001$, $r = .85$) and persisted after baseline differences were controlled. The advantage of AR appeared on five of the six indicators, especially inferring and comparing. These findings suggest that AR is most beneficial when it serves as an investigative tool in problem-based learning. Elementary teachers are encouraged to integrate AR into the investigation phase of PBL for abstract and spatial science content. Doing so requires enough Android devices for at least one per group, reliable internet access to load the AR content, printed markers, and teacher training in both the application and PBL facilitation. In schools in eastern Indonesia, where devices and connectivity are often limited, the shared-device group format used here offers a practical starting point. Further research should involve more schools, test long-term retention, and measure students' cognitive load directly.

REFERENCES

Ahmad, Z., Ahmad, H., & Rahman, Z. A. (2022). Penggunaan media pembelajaran augmented reality berbantuan Assemblr Edu untuk meningkatkan hasil belajar siswa SMA Negeri 5 Kota Ternate. *Jurnal Ilmiah Wahana Pendidikan, 8*(3), 514–521. https://doi.org/10.5281/zenodo.7421774

EDUKASI
Published:
Faculty of Teacher Training and Education
Universitas Khairun

e-ISSN 2715-8551

DOI: http://dx.doi.org/10.33387/j.edu.v24i2.13088
https://ejournal.unkhair.ac.id/index.php/edu
(Accredited Sinta Rank 3)

Index by:

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

Amelia, F., Hasibuan, A. U. H., Simamora, E. C., Ermawan, M. H., Siregar, N. S., Mirzah, N., & Gaol, S. I. L. (2025). Miskonsepsi siswa sekolah dasar tentang konsep IPA: Studi kasus di kelas VI SDN 024755. *Jurnal Bintang Pendidikan Indonesia, 3*(2), 207–217. https://doi.org/10.55606/jubpi.v3i2.3825

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Arici, F., & Yilmaz, R. M. (2023). An examination of the effectiveness of problem-based learning method supported by augmented reality in science education. *Journal of Computer Assisted Learning, 39*(2), 446–476. https://doi.org/10.1111/jcal.12752

Azuma, R. T. (1997). A survey of augmented reality. *Presence: Teleoperators and Virtual Environments, 6*(4), 355–385. https://doi.org/10.1162/pres.1997.6.4.355

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin, 114*(3), 494–509. https://doi.org/10.1037/0033-2909.114.3.494

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Faria, A. (2024). Augmented reality and teaching strategies in the study of volcanism in elementary and secondary schools. *Journal of New Approaches in Educational Research, 13*, Article 18. https://doi.org/10.1007/s44322-024-00018-5

Ferrari, E., Teijón, P. H., & Ruiz, C. (2024). Unlocking the cosmos: Evaluating the efficacy of augmented reality in secondary education astronomy instruction. *Journal of New Approaches in Educational Research, 13*, Article 8. https://doi.org/10.1007/s44322-024-00007-8

Fitriyani, F., Atmojo, I. R. W., & Yamtinah, S. (2025). Analysis of student misconceptions on the solar system material using the four-tier test instrument. *EduHumaniora: Jurnal Pendidikan Dasar, 17*(1), 11–22. https://doi.org/10.17509/eh.v17i1.69312

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141*(1), 2–18. https://doi.org/10.1037/a0024338

Garzón, J., & Acevedo, J. (2019). Meta-analysis of the impact of augmented reality on students' learning gains. *Educational Research Review, 27*, 244–260. https://doi.org/10.1016/j.edurev.2019.04.001

Gibran, M. A., & Setyasto, N. (2025). Effectiveness of augmented reality-assisted problem-based learning model on science learning outcomes. *Jurnal Penelitian Pendidikan IPA, 11*(2), 1093–1102. https://doi.org/10.29303/jppipa.v11i2.10319

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

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70. https://www.jstor.org/stable/4615733

Ibáñez, M.-B., & Delgado-Kloos, C. (2018). Augmented reality for STEM learning: A systematic review. *Computers & Education, 123*, 109–123. https://doi.org/10.1016/j.compedu.2018.05.002

Limpinan, P., Pichetphon, W., Yuensook, T., & Jantakoon, T. (2026). An experimental comparative study of virtual reality and augmented reality for teaching solar system in primary education. *Journal of Technology and Science Education, 16*(1), 262–274. https://doi.org/10.3926/jotse.3684

e-ISSN 2715-8551
DOI: http://dx.doi.org/10.33387/j.edu.v24i2.13088
https://ejournal.unkhair.ac.id/index.php/edu
(Accredited Sinta Rank 3)
Index by:

Mayer, R. E. (2024). The past, present, and future of the cognitive theory of multimedia learning. *Educational Psychology Review, 36*(1), Article 8. https://doi.org/10.1007/s10648-023-09842-1

Putri, F. E., Hermita, N., Zufriady, Z., & Wijaya, T. T. (2021). Identifying sixth grade students' misconceptions on solar system. *Journal of Teaching and Learning in Elementary Education, 4*(1), 65–76. https://doi.org/10.33578/jtlee.v4i1.7857

Quade, D. (1967). Rank analysis of covariance. *Journal of the American Statistical Association, 62*(320), 1187–1200. https://doi.org/10.1080/01621459.1967.10500925

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

Tran, T. N. A. (2024). Enhancing astronomy skills: The role of mobile augmented reality in Vietnamese middle schools. *International Journal of Interactive Mobile Technologies, 18*(22), 174–189. https://doi.org/10.3991/ijim.v18i22.51997

Uluçınar, U. (2023). The effect of problem-based learning in science education on academic achievement: A meta-analytical study. *Science Education International, 34*(2), 72–85. https://doi.org/10.33828/sei.v34.i2.1

Wen, Y., Wu, L., He, S., Ng, N. H.-E., Teo, B. C., Looi, C. K., & Cai, Y. (2023). Integrating augmented reality into inquiry-based learning approach in primary science classrooms. *Educational Technology Research and Development, 71*(4), 1631–1651. https://doi.org/10.1007/s11423-023-10235-y

Zufahmi, Z. (2025). Augmented reality in science learning: A systematic literature review. *Jurnal Pendidikan Biologi Indonesia, 11*(1), 274–291. https://doi.org/10.22219/jpbi.v11i1.38570