



Comparing Op-Board and Color-Coded Media for Third-Grade Pupils' Conceptual Understanding of Mixed Arithmetic Operations

Lutfiah Salsabilla, Fatkhul Arifin*

Universitas Islam Negeri Syarif Hidayatullah Jakarta, Indonesia

*Corresponding author's email: fatkhul_arf@uinjkt.ac.id

Abstract

Conceptual understanding of mixed arithmetic operations requires elementary pupils to connect problem information, operation symbols, calculation order, and mathematical explanations, yet comparative evidence on the instructional value of differently structured visual media remains limited. This study aims to compare third-grade pupils' conceptual understanding following instruction with Op-Board, a structured representational scaffold, and color-coded option media. The study employed a quasi-experimental pretest–posttest nonequivalent-group design involving 54 pupils in two intact classes (27 per class) at an Islamic elementary school in Indonesia. Over four instructional meetings, pupils completed a nine-item essay assessment covering concept restatement, information classification, and mathematical representation (reported Cronbach's $\alpha = .788$). Data were summarized descriptively, and posttest scores were compared using an independent-samples t-test and standardized effect size. The findings showed that mean scores increased from 56.30 to 90.52 in the Op-Board class and from 58.41 to 73.63 in the comparison class. The posttest difference favored Op-Board by 16.889 points, $t(52) = 8.730$, $p < .001$, 95% CI [13.007, 20.771], Cohen's $d = 2.38$. However, one intact class per condition and the absence of baseline adjustment preclude attributing this difference solely to the media. These findings contribute preliminary comparative evidence about representational organization, with implications for elementary mathematics instructional design and more rigorous classroom-level evaluations.

ARTICLE HISTORY

Received:

6 May 2026

Accepted:

19 September 2026

Published:

30 September 2026

KEYWORDS

representational
scaffolding;
conceptual
understanding;
mixed arithmetic
operations;
elementary
mathematics; visual
instructional media



1. Introduction

Elementary mathematics learning requires children to explain not only what an answer is but why its calculation is mathematically defensible. Mixed whole-number operations are a demanding setting for this requirement because pupils must distinguish the roles of addition, subtraction, multiplication, and division while attending to parentheses and precedence. A correct numerical answer may be produced through a remembered rule without revealing how a pupil



understands the structure of the expression. Therefore, Conceptual and procedural knowledge have been theorized as reciprocally developing rather than as separate or strictly sequential achievements (Rittle-Johnson et al., 2015). Longitudinal evidence on the understanding of arithmetic principles likewise treats conceptual relations as a meaningful learning outcome beyond routine performance (Göbel et al., 2026). Consequently, the educational problem is to make pupils' operation choices and reasoning accessible for learning and assessment, not merely to increase calculation speed.

When arithmetic is embedded in a word problem, successful performance requires several decisions that are cognitively and linguistically distinct. Pupils must identify the given quantities and the requested result, recognize what the quantities represent, translate their relationships into appropriate operation symbols, and execute the resulting expression in a justified sequence. A systematic review of elementary word-problem tasks found that linguistic features and the number or structure of operations are associated with task difficulty (Vessonen et al., 2024). Research on the visual perception of algebraic expressions shows that order-of-operations errors may also reflect how expressions are perceived, although that study focused on older students rather than third graders (Bye et al., 2024). Indonesian research documents related difficulties among elementary pupils dealing with whole-number operations and word problems (Nengsih & Pujiastuti, 2021). These distinct demands explain why seeing the quantities in a problem need not be equivalent to organizing them mathematically.

The nature of children's arithmetic reasoning also matters when procedures become more complex. Research on arithmetic strategy selection differentiates domain-general processes from arithmetic-specific knowledge, suggesting that calculation choices cannot be read solely from final accuracy (Eaves et al., 2022). Studies of strategy flexibility further distinguish adaptable mathematical reasoning from routine use of familiar algorithms (Jóelsdóttir et al., 2024). Experimental work involving finger-based learning illustrates that accessible representations may support early arithmetic in some instructional contexts, but its findings do not directly establish benefits for mixed-operation word problems in Grade III (Frey et al., 2024). Similarly, research tracing relationships from whole numbers to fractions and word problems concerns an older cohort and offers background rather than direct evidence for the present intervention (Xu et al., 2024). These studies collectively motivate attention to the links among task comprehension, operation selection, and expression structure.

External representations offer one route for making these links explicit, but their instructional function must be specified. Ainsworth's (2006) DeFT framework explains that multiple representations may complement information, constrain misinterpretation, and enable learners to construct connections among forms. A

meta-analysis of concrete manipulatives found that benefits depend on instructional conditions rather than on the mere presence of physical objects (Carbonneau et al., 2013). The account of concreteness fading adds that tangible, pictorial, and symbolic forms need to be connected progressively (Fyfe et al., 2014). Subsequent reviews of physical manipulatives and concrete–representational–abstract mathematics instruction reinforce the importance of how representations are used and linked in teaching (Byrne et al., 2023; Ebner et al., 2025). A recent review of primary teachers’ physical and virtual manipulative practices likewise underscores implementation as part of what makes a representation pedagogically meaningful (Ochugboju & Díez-Palomar, 2025).

Recent research refines this position by separating visual presentation from the cognitive and explanatory work stimulated by the visual resource. A systematic review of pictures accompanying mathematical word problems demonstrates that their usefulness depends on task and learner characteristics, rather than following from illustration alone (Vessonen et al., 2026). A meta-analysis of word-problem interventions similarly identifies implementation fidelity, dosage, outcome measures, and study quality as important sources of variation across findings (Vessonen et al., 2025). Explicitly prompting students to explain their mathematical ideas is also supported as an instructional consideration, although the quality of explanation cannot be inferred merely from a completed worksheet (Rittle-Johnson, 2024). Research on meaning-related language cautions against superficial translation from words to symbols that bypasses relational structure (Hankeln & Prediger, 2025). Thus, a theoretically informative comparison should ask what sequence of reasoning a medium supports, rather than whether pupils encountered something colorful or concrete.

In Indonesian elementary education, recent work has investigated diverse ways of improving mathematical learning, including concrete–pictorial–abstract approaches, interactive bulletin boards, and artificial-intelligence-assisted problem-based learning (Suryaningsih et al., 2025; Setiawan et al., 2025; Noviyana et al., 2025). A synthesis of technology-assisted mathematics instruction also identifies potential opportunities for numeracy learning, but it concerns a broader family of interventions than non-digital operation boards (Ariani et al., 2024). These investigations contribute a relevant pedagogical context without answering whether two active visual media with different degrees of explicit reasoning structure produce different classroom score profiles for mixed arithmetic. Furthermore, language-responsive materials can support mathematical learning in heterogeneous classrooms, yet results from fraction instruction do not establish the effectiveness of a particular sequence scaffold for Grade III mixed operations (Lenz et al., 2024). Therefore, the unresolved issue is more specific than whether visual media are generally beneficial: it concerns the organization of mathematical reasoning within the visual support.

The present study compared two existing Grade III classes using different active visual resources. Op-Board combined a space for classifying known and requested information, an operation-symbol selection area, a visible calculation-priority sequence, and a worked-solution pathway; the color-coded option medium highlighted operations and priorities without the same integrated record of reasoning. The study asked whether their overall essay-based conceptual understanding scores differed after four meetings on the same mathematical topic. Its measurement was organized around restating concepts, classifying problem information, and expressing mathematical relationships through representations, but it reported total scores rather than separately scored indicators. The central contribution is a context-specific comparison between structured representational coordination and simpler operational cues, rather than between visual and non-visual materials.

2. Methods

The study used a quantitative quasi-experimental, pretest–posttest nonequivalent-groups design during the 2024/2025 academic year at MI Jam’iyyatul Khair, Indonesia. The school’s two existing third-grade classes constituted the observed sample ($N = 54$). Class III A ($n = 27$) received instruction with Op-Board, while Class III B ($n = 27$) used color-coded option media. The pupils were not individually randomized, and there was no random allocation of a larger set of classrooms. Accordingly, classroom membership and instructional medium coincide exactly in this study: there is one intact class per condition. Although the design allows the reported score profiles to be compared, it does not create independent classroom-level treatment replication.

The instructional plan specified the same mixed whole-number arithmetic topic, learning objectives, teacher, number of teaching meetings, planned instructional duration, and assessment instrument in both classes. The pretest was administered before the media-specific instruction to describe each class’s initial assessment results; it was not employed as a covariate in the originally reported inferential model. Keeping planned instruction comparable reduces some obvious differences between conditions, but without independent replication and recorded fidelity scores, the study cannot disentangle the contribution of the medium from all class-specific experiences. The resulting analytical unit in the reported t-test was the pupil score; the actual instructional allocation unit was the class.

Op-Board was designed as a visual-concrete working surface with four coordinated functions: recording the information supplied and requested by a problem, selecting operation symbols, sequencing computations according to priority rules, and presenting successive calculation steps together with an explanation. Its purpose was to keep the movement from verbal quantities to

symbolic relations visible to pupils and the teacher. In the comparison class, color-coded option media were used to distinguish operation types or priorities through color cues; the comparison medium did not provide equivalent dedicated spaces for the full reasoning sequence. Both classes completed comparable worksheets and card-based tasks and had opportunities for group interaction, discussion, and presentation. Consequently, the planned contrast was between two active forms of visual support, not between visual instruction and a no-media control.

The four-meeting Op-Board sequence moved from modeling to increasingly pupil-directed problem work. During Meeting 1, the teacher introduced precedence through a contextual task, modeled the identification of relevant information and selection of operations, and asked groups to explain their choices. Meeting 2 used collaborative worksheet problems and shared presentations, with the board available as a guide to solution order. During Meeting 3, groups completed timed question-card activities followed by a reflection on operational priorities. In Meeting 4, each group constructed three mixed-operation problems, exchanged them with another group, worked through the received problems, explained selected solutions, and completed individual written work. Similar types of group tasks and time allocations were planned for the color-coded comparison class; the detailed sequence of activities is best documented for Op-Board. Observation sheets recorded the conduct of instructional steps and media use, but no numerical implementation-fidelity index was supplied.

The primary outcome was the overall score on an essay-based mathematical conceptual-understanding test. Ten questions were initially written to cover three domains: explaining or restating arithmetic concepts and precedence, identifying and classifying information according to problem characteristics, and presenting mathematical relations through symbolic expressions or ordered solutions. A mathematics-education expert reviewed the items for mathematical correctness, indicator alignment, clarity of wording, and appropriateness for Grade III. The revised questions were piloted among pupils one grade above the target group. The source study reports item-total correlation testing against a critical r of .374 at the .05 significance level, leading to the retention of nine items and the removal of one. The resulting reported Cronbach's alpha was .788, and the same nine questions were administered at pretest and posttest.

The source documents do not specify the size of the pilot sample, complete item coefficients, the analytic rubric, item weights, or inter-rater agreement procedures. Therefore, the available evidence supports a comparison of the reported total scores but not a reproducible reconstruction of how essay responses became numerical scores or a numerical comparison of the three

assessment domains. Observation records and pupils' written work were used to describe the learning activities and illustrate relevant mathematical tasks; they were not subjected to a separately reported coding analysis and were not treated as an additional qualitative strand. Repeating the identical assessment items at both occasions also creates a possibility of test-format familiarity contributing to the observed change.

The source analysis was conducted using IBM SPSS Statistics 25. Descriptive summaries included the number of pupils, means, standard deviations, minima, maxima, and variances for each class at both occasions. The Kolmogorov–Smirnov and Shapiro–Wilk procedures screened distributions, while reported Levene tests assessed equality of variances. The principal inferential result was a two-sided independent-samples t-test comparing posttest total scores using the equal-variances-assumed output row and a nominal α of .05, accompanied by a reported mean difference, standard error, and 95% confidence interval. The pooled posttest standard deviation was used for Cohen's d , with Hedges' g and eta squared reported as complementary posttest contrast indices.

Absolute mean-score change was calculated as the posttest minus pretest mean within each class. The normalized gain presented in the source is based on the group means, specifically $(\text{posttest group means} - \text{pretest group mean}) / (100 - \text{pretest group mean})$, using 100 as the maximum possible score. Neither a paired-score analysis nor a baseline-adjusted model is available. The t-test treats the 54 pupil scores as independent observations even though instructional media were delivered within two intact classrooms. Its p-value, interval, and standardized difference consequently describe the observed pupil-level contrast under that model, but cannot provide classroom-replicated causal evidence. In particular, the analysis cannot estimate an independent variation in medium effects across multiple classrooms.

3. Results

The study examined differences in mathematical conceptual-understanding outcomes between two Grade III classes taught mixed arithmetic operations with different visual learning media. Class III A (27 pupils) used Op-Board, which organizes the stages of problem solving, whereas Class III B (27 pupils) used color-coded option media that highlight operation types and priorities. Both groups completed the same nine-item essay assessment before instruction and after four teaching meetings. The assessment addressed three domains—concept restatement, information classification, and mathematical representation, but the numerical dataset available for reporting consists of overall test scores, not separate scores for these domains. Therefore, the results progress from initial score profiles and descriptive changes to illustrative task

evidence, assumption checks, and the reported posttest contrast. All reported comparisons refer to these two observed classes.

3.1 Initial and final score distributions differed between classes

The initial comparison establishes the two classes' observed starting points before either medium was used. For each group, the pretest and posttest were summarized using the number of pupils, mean score, standard deviation, score range, and variance. Examining both central tendency and dispersion is necessary because a single mean cannot show whether performance was concentrated within a narrow band or spread across markedly different score levels. The summary statistics for the four group-by-occasion combinations are presented in Table 1.

Table 1.

Descriptive distribution of conceptual-understanding scores

Class	Assessment	n	Mean	SD	Observed range	Variance
Op-Board	Pretest	27	56.30	7.194	44–67	51.755
Op-Board	Posttest	27	90.52	6.154	81–100	37.875
Color-coded option	Pretest	27	58.41	5.672	48–67	32.174
Color-coded option	Posttest	27	73.63	7.948	63–93	63.165

Note. SD = standard deviation. Each class included 27 pupils on both assessment occasions. Scores were reported on a scale with an upper bound of 100.

Before instruction, the mean in the color-coded option class (58.41) was slightly higher than in the Op-Board class (56.30), a difference of 2.11 points. Both classes had the same highest observed pretest score of 67, although their minima differed (44 for Op-Board and 48 for the comparison class). The Op-Board class showed somewhat greater initial dispersion, with an SD of 7.194 versus 5.672. Thus, the initial score distributions were not identical, even though their averages occupied a relatively similar range. These baseline statistics describe the groups before teaching; they are not evidence that the classes were equivalent on every relevant characteristic.

After the four meetings, the ordering of the class means had reversed. The Op-Board class reached a posttest mean of 90.52, whereas the comparison class reached 73.63. This 16.89-point difference was accompanied by different observed score intervals: 81–100 in Op-Board and 63–93 in the comparison group. The lowest recorded Op-Board posttest score (81) exceeded the highest recorded pretest score (67) in either class. This is a comparison of reported marginal score ranges, rather than a reconstruction of individual pupil progress. In the comparison class, the posttest range of 63–93 overlapped with its initial range of 48–67.

Dispersion also changed in different directions. Op-Board posttest scores had a smaller standard deviation than at pretest (6.154 versus 7.194), while the color-coded option class had a larger posttest standard deviation (7.948 versus 5.672). The variance entries show the same directional pattern, decreasing from 51.755 to 37.875 in Op-Board and increasing from 32.174 to 63.165 in the comparison class. Taken together, the descriptive evidence shows similar-range initial class means followed by a higher, more concentrated posttest distribution in the Op-Board class and a lower, more dispersed distribution in the comparator.

3.2 Both classes increased their mean scores, but the descriptive gains differed

Because the same conceptual-understanding assessment was administered before and after instruction, changes in group means can be described directly. For each class, the absolute mean-score increase was calculated by subtracting the pretest mean from the posttest mean. In addition, mean-based normalized gain was calculated as $(\text{posttest mean} - \text{pretest mean}) / (100 - \text{pretest mean})$, expressing the increase relative to the remaining possible mean-score improvement. These two measures describe change from different perspectives: the absolute increase uses test-score points, while normalized gain reflects the proportion of the mean-level improvement opportunity realized. Table 2 presents the calculations for both classes.

Table 2.

Mean-score changes and group-mean normalized gains

Instructional class	Pretest mean	Posttest mean	Increase (points)	Normalized gain	Reported category
Op-Board	56.30	90.52	34.22	0.78	High
Color-coded option	58.41	73.63	15.22	0.37	Medium

The Op-Board mean increased by 34.22 points, from 56.30 to 90.52. Therefore, the comparison mean increased by 15.22 points, from 58.41 to 73.63. The observed difference between these two mean increases was 19.00 points. This change contrast is distinct from the 16.89-point posttest difference: the former accounts arithmetically for the fact that the Op-Board class started 2.11 points below the comparison class, whereas the latter compares only final mean scores. Both descriptions identify a larger upward shift in the Op-Board class, but the underlying pupil-level paired-score data were not available for estimating the variability or statistical uncertainty of individual gains.

The normalized-gain values show a parallel pattern. The mean-based gain was 0.78 for the Op-Board class and 0.37 for the color-coded option class, a difference of 0.41. Expressed in the categorical scheme reported in the original analysis, these corresponded to high and medium gain, respectively. Such categories summarize group-level change but should not be treated as evidence

that every pupil experienced the same amount of improvement. Nor is a difference between these gain indices a replacement for a baseline-adjusted estimate of the instructional contrast.

Figure 1 displays the change in group means without implying that individual pupil trajectories were uniform. The two lines begin at nearby pretest levels but separate substantially at posttest, mirroring the pattern in Table 2.

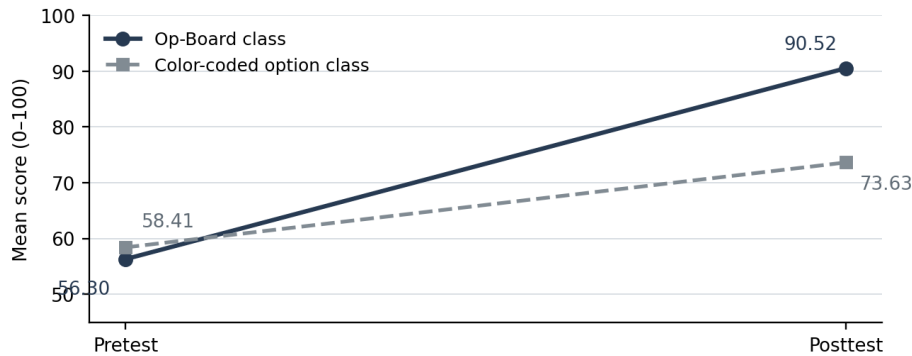


Fig. 1. Pretest and posttest group means.

*Notes: Lines connect group averages across assessment occasions;
They do not represent individual pupil trajectories.*

The shared empirical pattern is a positive mean-score change in both instructional conditions, with a larger absolute and normalized increase in the Op-Board class. Because scores were summarized at two occasions and individual paired differences were not supplied, this subsection establishes the direction and magnitude of observed class-level changes but does not test their causal origin.

3.3 Documented assessment tasks illustrate the three domains of conceptual understanding

The nine-item essay assessment was designed around three indicators: restating arithmetic concepts, classifying information according to problem requirements, and representing mathematical relationships through expressions and written solutions. The earlier study record also describes selected pupil work from tasks addressing these domains. This material helps clarify what the total scores were intended to capture beyond a correct final calculation. Unlike the numerical results above, however, these classroom-work descriptions were not accompanied by item-level scores, coded response frequencies, or independently analyzed qualitative themes. Table 3 presents them as illustrations of assessment content and documented answer characteristics, not as three separately tested outcomes.

Table 3.
Selected task contexts and described response features across assessment domains

Assessment domain	Illustrative assessment demand	Description available in the documented study account
Restating concepts	Explain which arithmetic operation should be performed first in a mixed expression without parentheses.	Selected Op-Board responses were described as offering more complete order-of-operations explanations; some comparison responses gave an incorrect restatement.
Classifying information	Identify quantities of red and blue marbles, marbles given away, and the remaining quantity before calculation.	Selected Op-Board work was described as organizing the given information first; some comparison work proceeded directly to computation.
Mathematical representation	Translate a word problem about pencils for male and female pupils into an expression and a sequence of solution steps.	Selected Op-Board work was described as including operations and steps; some comparison answers lacked an explicit representation or explanation.

Note. The entries synthesize task and response descriptions included in the study documentation. No counts, quotations, domain averages, or between-domain inferential tests were available.

These examples indicate that the assessment was intended to extend beyond computation alone. The first type of item elicited an explanation of operation precedence, the second required pupils to distinguish given from requested information in a contextual situation, and the third asked pupils to translate written information into symbolic mathematical work. The available account describes differences in selected worked answers between the two classrooms; it does not provide how many pupils displayed each characteristic or whether the selected examples represented the broader distribution of responses. Accordingly, the examples are consistent with the stated assessment domains but cannot be used to quantify or statistically verify separate improvements in restatement, classification, or representation. Therefore, the pattern established by this documentary evidence is one of assessment coverage rather than indicator-level outcome separation. The available record links the overall essay score to three concrete kinds of mathematical work, while the principal numerical finding remains the total-score contrast between classes.

3.4 Distributional and variance checks supported the reported test specification only at the pupil-score level

Before the reported comparison of posttest means, the original analysis assessed the distributions of scores using the Kolmogorov–Smirnov and Shapiro–Wilk tests for both classes at both assessment times. It also reported Levene's tests of equality of variances between groups. These checks concern properties of the observed pupil-score distributions under the selected statistical model. Their results are summarized separately in Tables 4 and 5 so that the

normality and variance information can be read without conflating distinct assumptions.

Table 4.

Reported normality-test significance values

Occasion	Instructional class	Kolmogorov–Smirnov p	Shapiro–Wilk p
Pretest	Op-Board	.121	.096
Pretest	Color-coded option	.200	.113
Posttest	Op-Board	.200	.058
Posttest	Color-coded option	.133	.131

Note. Values reproduce the significance entries in the original reported normality table. A printed value of .200 is retained as displayed in the source.

All eight normality-test p-values exceeded the stated .05 criterion. For Shapiro–Wilk, the reported p-values were .096 and .058 in the Op-Board at pretest and posttest, respectively, and .113 and .131 in the comparison group. The smallest was .058 for the Op-Board posttest. For Kolmogorov–Smirnov, the corresponding values ranged from .121 to .200. Thus, none of the reported tests identified a departure from normality at the conventional .05 cutoff. This pattern means that the null hypothesis of normality was not rejected by the tests; it does not demonstrate that the four underlying distributions were exactly normal.

Table 5.

Reported Levene tests of equality of variances

Assessment/source output	Levene F	df1	df2	p
Pretest, separate homogeneity table	1.140	1	52	.291
Posttest, separate homogeneity table	1.832	1	52	.182
Posttest, t-test output	0.892	1	52	.349

The separately presented Levene tables reported $p = .291$ for pretest and $p = .182$ for posttest; the Levene entry shown alongside the independent-samples t-test instead gave $p = .349$ for posttest. Although the two posttest results differ numerically, each is above .05 and neither indicates rejection of the equal-variance assumption at the reported threshold. The original inferential table used the row labeled “equal variances assumed,” which is retained in the posttest analysis reported below. Importantly, neither Levene’s test nor the distributional tests addresses dependence among pupils who belong to the same classroom. In summary, the available diagnostic outputs did not indicate departures from normality or unequal variances at the selected significance level. Those results explain the pupil-level t-test specification reported in the study, but do not remove the study-design limitation that each medium was used in just one intact classroom.

3.5 The reported posttest comparison showed a large separation in total scores

The principal statistical test addressed whether the mean conceptual-understanding score at posttest differed between the Op-Board and color-coded option classes. The reported analysis was a two-sided independent-samples t-test using the equal-variances-assumed row with 52 degrees of freedom. Beyond the p-value, the original output provided the mean difference, standard error, and confidence interval, allowing the location and magnitude of the observed contrast to be described. Standardized indices were also available for characterizing that contrast relative to the within-class posttest spread. Table 6 presents the full set of reported test statistics and magnitude descriptors.

Table 6.
Independent-samples comparison of posttest total scores

Inferential statistic	Reported result
Contrast	Op-Board minus color-coded option
Difference between posttest means	16.889 points
Standard error of the mean difference	1.934
Independent-samples test	$t(52) = 8.730$
Two-sided significance	$p < .001$
Reported 95% confidence interval	[13.007, 20.771]

Note. The results reproduce the equal-variances-assumed row in the source output. The software entry .000 is conventionally represented as $p < .001$.

The mean posttest difference was 16.889 points in favor of the Op-Board class. The reported test yielded $t(52) = 8.730$ and a two-sided p-value below .001. Its 95% confidence interval extended from 13.007 to 20.771 score points, with a standard error of 1.934. Every value in that interval was positive, so the reported pupil-level test consistently favored the Op-Board class under its assumptions. The confidence interval is an interval for the difference in the model used, not a guarantee that a comparable difference would recur after independently allocating media across additional classrooms.

Table 7.
Standardized magnitude of the observed posttest separation

Descriptive index	Reported value
Pooled posttest standard deviation	7.108
Cohen's d	2.38
Hedges' g	2.34
Eta squared (η^2)	.594

Note. These indices summarize the observed pupil-score contrast; they are not independent causal treatment-effect estimates.

The standardized separation was also substantial. Dividing the posttest mean difference by the pooled posttest SD of 7.108 gave the reported Cohen's d of approximately 2.38, and the small-sample-corrected Hedges' g was 2.34. Both statistics describe a large gap relative to the variability of the two observed posttest distributions. The reported eta squared value of .594 similarly summarizes the association between group membership and posttest scores in the fitted individual-level comparison. It does not mean that Op-Board independently caused 59.4% of pupils' learning or that an identical magnitude should be expected in other schools.

The statistical results and the descriptive changes point in the same observed direction: the Op-Board class finished with a higher average score, a narrower range of observed posttest scores, and a larger mean-level increase than the comparison class. Nevertheless, each instructional medium was delivered to one pre-existing class, so pupils' classroom membership and medium exposure cannot be separated in the available design. Consequently, the t -test and effect-size estimates summarize these two classrooms rather than demonstrating a classroom-replicated causal effect. The empirical conclusion is a clear and substantial observed posttest difference in overall conceptual-understanding scores, accompanied by distinct patterns of descriptive change.

4. Discussion

The principal observation is a clear divergence in immediate conceptual-understanding total scores between two existing classrooms that began with relatively nearby mean scores. Both classes experienced higher average scores after the four lessons, yet the class using an integrated reasoning pathway recorded the larger descriptive change and the higher posttest level. The interpretive question is not whether seeing a board is better than seeing no board, since both groups used visual resources, but whether an instructional resource that makes relationships and reasoning steps explicit could be educationally distinctive. This study offers comparative classroom evidence compatible with that proposition, while the allocation of one classroom per medium leaves causal attribution unresolved. The scholarly significance thus lies in defining a testable distinction between visual cueing and representational coordination, rather than in declaring Op-Board a universally superior product.

The first plausible explanatory process is the external organization of semantic-to-symbolic mapping. A word problem requires pupils to decide what quantities mean before determining an operation; otherwise, familiar numbers and symbols can trigger premature calculation without a defensible mathematical model. The Op-Board sequence begins with known and requested information, then makes the operation choice and calculation order explicit on the same working surface. In the DeFT framework, the benefit of combined representations

comes from functional connections and restrictions on interpretation, not simply from display density (Ainsworth, 2006). A recent systematic review of task characteristics in elementary word-problem solving similarly shows that the structure and linguistic features of tasks alter their difficulty (Vessonen et al., 2024). Schema-instruction research addresses the value of making underlying mathematical relations recognizable, although the present board was not tested as a formal schema-based intervention (Hardy et al., 2026). Therefore, the observed classroom difference is consistent with, but does not itself demonstrate, an account centered on externalized problem structure.

A second plausible pathway concerns the visibility of pupils' mathematical explanations. A scaffold that asks pupils to record the reason for choosing an operation may help separate a justified calculation from an answer obtained by rote execution, creating an opportunity for the teacher to assess the quality of reasoning. The rationale accords with research encouraging mathematical explanation and with analyses emphasizing the pedagogical construction of reasoning as an object of classroom work (Rittle-Johnson, 2024; Alderton & Pratt, 2025). Work on linguistic meaning in mathematics shows the importance of relational language rather than superficial word-to-symbol substitution (Hankeln & Prediger, 2025), while investigations of prospective teachers' dialogic moves reveal that eliciting a response is not identical to extending its mathematical reasoning (Shure & Liljedahl, 2024). Neither teacher–pupil dialogue nor independently coded explanations were measured in this study. Accordingly, explanatory accountability remains a theoretically grounded account of what the board could afford, rather than a mediator established by the data.

The distinction between conceptual and procedural learning is especially important for interpreting the assessment. Restating a precedence rule addresses the pupil's ability to articulate a mathematical principle; choosing and performing operations requires procedural execution; classifying quantities and translating verbal conditions into expressions link these levels of knowledge. Reciprocal relations between conceptual and procedural knowledge have been documented and theorized in mathematics education (Rittle-Johnson et al., 2015), while more recent longitudinal analysis investigates the development of arithmetic-principle understanding across primary years (Göbel et al., 2026). Research on flexible multidigit strategies similarly cautions that reliable routine performance and adaptable expertise are not interchangeable (Jóelsdóttir et al., 2024). Work on primary pupils' generalization with block-based tasks further illustrates the importance of reasoning about mathematical relationships (Wilkie & Hopkins, 2024). Op-Board was deliberately arranged to keep conceptual judgments next to written computational work. However, the reported outcome was an aggregate essay score: without scored subdomains, the present evidence cannot say whether explaining rules, classifying information, or representing relations improved most.

The documented task examples provide a limited but useful window into the construct behind that aggregate score. In one assessment item, a pupil was expected to explain which operation takes precedence in an expression without parentheses; another problem required separating red and blue marble quantities, marbles given away, and the requested remainder. A further example involved pencils for pupils and asked for a symbolic representation and the sequence of calculations. These contexts explain why the assessment was designed to observe understanding beyond the production of a final number. The earlier classroom account describes selected Op-Board work as providing clearer information organization and explanations than some comparator responses. Yet no response frequencies, item-level means, transcripts, or coded criteria accompany those observations. Therefore, it would be an overstatement to report three significant indicator-specific effects, and the examples are better understood as illustrations of intended assessment coverage.

The intervention can also be situated within the wider literature on concrete and visual mathematical learning. Meta-analytic evidence suggests that concrete manipulatives can support mathematics outcomes under appropriately guided conditions (Carbonneau et al., 2013), while concreteness-fading research describes the need to connect tangible and symbolic forms rather than treating concrete objects as educational ends in themselves (Fyfe et al., 2014). Laski et al. (2015) also emphasize connecting manipulative activity with mathematical meaning. Reviews of physical manipulatives and concrete–representational–abstract instruction add that implementation and representational transitions are consequential features of instruction (Byrne et al., 2023; Ebner et al., 2025). Studies of blended concrete and virtual manipulatives and primary teachers' physical and virtual material use provide further examples of varied delivery conditions, not a direct replication of this board comparison (Siller & Ahmad, 2024; Ochugboju & Díez-Palomar, 2025). Research on finger-supported first-grade arithmetic and broader learning strategy variation likewise underlines that mechanisms and age groups should not be conflated across interventions (Frey et al., 2024; Eaves et al., 2022). Therefore, the current study's distinction is functional: the board linked representation, decision, and justification in a visible sequence.

The active comparison condition is a meaningful feature of the design. Since color-coded option media also visually highlight operations and precedence, the comparison does not establish that visual resources are categorically better than text-only lessons, nor does it imply that color cues have no pedagogical value. Other approaches in elementary mathematics—including concrete–pictorial–abstract numeracy lessons, bulletin-board supports for early number sense, and technology-assisted problem solving—have been studied with different outcomes and age groups (Suryaningsih et al., 2025; Setiawan et al., 2025; Noviyana et al., 2025). Research reviews of ICT-supported numeracy similarly concern a broader

instructional class than Op-Board (Ariani et al., 2024). Analyses of language-responsive materials show that instructional benefits can vary with students' prior knowledge or linguistic conditions (Lenz et al., 2024), and broader arithmetic knowledge studies trace dependencies across mathematical topics without testing the media examined here (Xu et al., 2024). Accordingly, a productive contribution of this comparison is the design distinction it makes available for future tests, not a general claim that structured boards always outperform alternatives.

A substantial numerical contrast should also be distinguished from a credible causal effect estimate. The observed Cohen's d of 2.38, Hedges' g of 2.34, and eta squared of .594 are large descriptors of the individual-level posttest difference; they do not neutralize selection, confounding, or common classroom influences. Because all 27 pupils within a condition shared the same classroom, there is no independent variation in the medium at the classroom level. Pupil-level assumptions checked by the Shapiro–Wilk and Levene tests cannot identify or correct classroom clustering, and baseline adjustment alone cannot manufacture the missing classroom replication. The descriptive score changes may also reflect instruction received, familiarity with repeated questions, teacher interaction, initial group characteristics, or combinations of these factors. This is why an interval derived from the original pupil-level t -test should not be treated as a generalizable confidence interval for the effect of adopting the medium across schools.

Measurement uncertainty further qualifies the interpretation. The nine retained questions were reportedly screened after expert review and a pilot, and Cronbach's alpha of .788 provides some internal-consistency information for the reported instrument. However, without the full scoring rubric, item weights, pilot sample details, or independent assessment of essay ratings, measurement reproducibility remains limited. The two non-matching Levene posttest entries in the source should not be silently merged: both are above .05, but they are distinct reported numerical outputs whose procedural relationship is unknown. These reporting constraints matter particularly for claims about the size and precision of effects, even though they do not erase the fact that the observed posttest means were separated by nearly 17 points. The immediate four-meeting timescale also means that delayed retention and transfer to unfamiliar problem formats remain outside the current evidence.

The practical implication is to use visual aids as prompts for a complete mathematical argument rather than solely as displays of operations. Teachers could ask pupils to list what is known and requested, state why an operation represents the relationship, order the calculations with reference to parentheses and precedence, and check the resulting answer using another representation when appropriate. Such a sequence is most defensible as a teaching design to

be investigated and refined, not as a proven mechanism of the score difference in this particular study. Further research should allocate multiple independent classrooms to each medium, use a documented essay-scoring rubric with rater checks, and examine item-specific evidence, worked solutions, and classroom explanations. Longer follow-up and unfamiliar transfer tasks would help test whether early advantages remain when prompts are reduced. In scholarly terms, the present study motivates closer examination of how the *organization* of representations interacts with opportunities to justify mathematical choices.

5. Conclusion

This classroom comparison found a marked immediate difference in mathematical conceptual-understanding total scores after four lessons on mixed arithmetic operations. The class using Op-Board moved from a pretest mean of 56.30 to 90.52, while the class using color-coded option media moved from 58.41 to 73.63; the observed posttest separation was 16.89 score points. Both classrooms improved on the reported aggregate assessment, but the magnitude and final distributions were different. Because the classes received two active forms of visual support, the central empirical distinction is not the presence versus absence of media. Instead, it is the contrast between an integrated visible solution pathway and more limited operation cues that merits closer examination in elementary mathematics.

The study contributes a precise, context-specific question to debates about multiple representations and elementary arithmetic: whether visual resources that coordinate what a problem says, which operations are selected, how calculations are ordered, and why steps are justified can support understanding differently from visual cues alone. The evidence supports the existence of an observed total-score contrast, not a causal explanation of that contrast or evidence of separately improved assessment domains. One classroom was used per medium, paired and baseline-adjusted analyses were not available, and details of essay scoring were incomplete. These conditions delimit rather than nullify the descriptive findings. Stronger classroom-randomized or replicated quasi-experimental designs, transparent assessment procedures, and process evidence are needed to establish when and for whom representational scaffolding offers a durable advantage.

Declarations

The authors got school permission, parental or guardian consent, and pupil assent. Authors have no competing interests. Grammarly was used to check language. No data generated by AI. The authors are responsible for verifying the scientific content, references, numerical results, and final declarations.

References

- Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction*, 16(3), 183–198. <https://doi.org/10.1016/j.learninstruc.2006.03.001>
- Alderton, J., & Pratt, N. (2025). The construction of mathematical reasoning as a pedagogical object. *Educational Studies in Mathematics*, 120(2), 299–315. <https://doi.org/10.1007/s10649-025-10420-1>
- Ariani, Y., Suparman, Helsa, Y., Zainil, M., & Rahmatina. (2024). ICT-based or -assisted mathematics learning and numerical literacy: A systematic review and meta-analysis. *International Journal of Information and Education Technology*, 14(3), 382–397. <https://doi.org/10.18178/ijiet.2024.14.3.2060>
- Bye, J. K., Chan, J. Y.-C., Closser, A. H., Lee, J.-E., Shaw, S. T., & Ottmar, E. R. (2024). Perceiving precedence: Order of operations errors are predicted by perception of equivalent expressions. *Journal of Numerical Cognition*, 10, Article e14103. <https://doi.org/10.5964/jnc.14103>
- Byrne, E. M., Jensen, H., Thomsen, B. S., & Ramchandani, P. G. (2023). Educational interventions involving physical manipulatives for improving children's learning and development: A scoping review. *Review of Education*, 11(2), Article e3400. <https://doi.org/10.1002/rev3.3400>
- Carbonneau, K. J., Marley, S. C., & Selig, J. P. (2013). A meta-analysis of the efficacy of teaching mathematics with concrete manipulatives. *Journal of Educational Psychology*, 105(2), 380–400. <https://doi.org/10.1037/a0031084>
- Eaves, J., Attridge, N., & Gilmore, C. (2022). The role of domain-general and domain-specific skills in the identification of arithmetic strategies. *Journal of Numerical Cognition*, 8(3), Article e7459. <https://doi.org/10.5964/jnc.7459>
- Ebner, S., MacDonald, M. K., Grekov, P., & Aspiranti, K. B. (2025). A meta-analytic review of the concrete-representational-abstract math approach. *Learning Disabilities Research & Practice*, 40(1), 31–42. <https://doi.org/10.1177/09388982241292299>
- Frey, M., Gashaj, V., Nuerk, H.-C., & Moeller, K. (2024). You can count on your fingers: Finger-based intervention improves first-graders' arithmetic learning. *Journal of Experimental Child Psychology*, 244, Article 105934. <https://doi.org/10.1016/j.jecp.2024.105934>
- Fyfe, E. R., McNeil, N. M., Son, J. Y., & Goldstone, R. L. (2014). Concreteness fading in mathematics and science instruction: A systematic review. *Educational Psychology Review*, 26(1), 9–25. <https://doi.org/10.1007/s10648-014-9249-3>

- Göbel, S. M., Landerl, K., & Lervåg, A. O. (2026). Longitudinal predictors of conceptual understanding of arithmetic principles. *Journal of Numerical Cognition*, 12, Article e19523. <https://doi.org/10.5964/jnc.19523>
- Hankeln, C., & Prediger, S. (2025). Language is essential for avoiding surface translations: Associations of students' spontaneous use of meaning-related phrases for explicating structures with conceptual understanding of multiplication. *Educational Studies in Mathematics*, 120, 57–79. <https://doi.org/10.1007/s10649-025-10414-z>
- Hardy, A. M., Powell, S. R., & Saglam-Ak, A. (2026). Implementing schema instruction to support young children with word problems: A systematic review. *Journal of Numerical Cognition*, 12, Article e17649. <https://doi.org/10.5964/jnc.17649>
- Jóelsdóttir, L. B., Sunde, P. B., Sunde, P., & Andrews, P. (2024). Routine and adaptive experts: Individual characteristics and their impact on multidigit arithmetic strategy flexibility and mathematics achievement. *Journal of Numerical Cognition*, 10, Article e14081. <https://doi.org/10.5964/jnc.14081>
- Laski, E. V., Jor'dan, J. R., Daoust, C., & Murray, A. K. (2015). What makes mathematics manipulatives effective? Lessons from cognitive science and Montessori education. *SAGE Open*, 5(2), Article 2158244015589588. <https://doi.org/10.1177/2158244015589588>
- Lenz, K., Obersteiner, A., & Wittmann, G. (2024). Who benefits most from language-responsive learning materials in mathematics? Investigating differential effects in heterogeneous classrooms. *Educational Studies in Mathematics*, 116, 185–211. <https://doi.org/10.1007/s10649-024-10321-9>
- Nengsih, G. A., & Pujiastuti, H. (2021). Analisis kesulitan dalam menyelesaikan soal materi operasi bilangan cacah siswa sekolah dasar. *JKPM (Jurnal Kajian Pendidikan Matematika)*, 6(2), 293–306. <https://doi.org/10.30998/jkpm.v6i2.9941>
- Noviyana, H., Rahmawati, F., Kirana, A. R., & Tanod, M. J. (2025). Enhancing elementary students' mathematical problem-solving skills through AI-assisted problem-based learning. *Journal of Integrated Elementary Education*, 5(2), 254–268. <https://doi.org/10.21580/jieed.v5i2.27576>
- Ochugboju, A. O., & Díez-Palomar, J. (2025). Systematic review on the use of virtual and physical manipulatives by primary school teachers. *Social Sciences & Humanities Open*, 12, Article 102188. <https://doi.org/10.1016/j.ssaho.2025.102188>
- Rittle-Johnson, B. (2024). Encouraging students to explain their ideas when learning mathematics: A psychological perspective. *The Journal of Mathematical Behavior*, 76, Article 101192. <https://doi.org/10.1016/j.jmathb.2024.101192>

- Rittle-Johnson, B., Schneider, M., & Star, J. R. (2015). Not a one-way street: Bidirectional relations between procedural and conceptual knowledge of mathematics. *Educational Psychology Review*, 27(4), 587–597. <https://doi.org/10.1007/s10648-015-9302-x>
- Setiawan, Y., Marwa, K., Borom, A., & Mamji, Y. (2025). The efficacy of the “Pojok Ceria” interactive bulletin board in enhancing early number sense and self-directed learning among first-grade students. *Journal of Integrated Elementary Education*, 5(2), 506–526. <https://doi.org/10.21580/jieed.v5i2.23895>
- Shure, V., & Liljedahl, P. (2024). The use of a scriptwriting task as a window into how prospective teachers envision teacher moves for supporting student reasoning. *Journal of Mathematics Teacher Education*, 27, 411–440. <https://doi.org/10.1007/s10857-023-09570-x>
- Siller, H.-S., & Ahmad, S. (2024). The effect of concrete and virtual manipulative blended instruction on mathematical achievement for elementary school students. *Canadian Journal of Science, Mathematics and Technology Education*, 24, 229–266. <https://doi.org/10.1007/s42330-024-00336-y>
- Suryaningsih, T., Nisa, K., Widiyanto, R., Latip, A. E., & Siron, Y. (2025). Enhancing elementary students’ numeracy skills through the concrete pictorial abstract (CPA) approach. *Journal of Integrated Elementary Education*, 5(1), 224–236. <https://doi.org/10.21580/jieed.v5i1.25383>
- Vessonen, T., Dahlberg, M., Hellstrand, H., Widlund, A., Korhonen, J., Aunio, P., & Laine, A. (2024). Task characteristics associated with mathematical word problem-solving performance among elementary school-aged children: A systematic review and meta-analysis. *Educational Psychology Review*, 36, Article 117. <https://doi.org/10.1007/s10648-024-09954-2>
- Vessonen, T., Hellstrand, H., Aunio, P., & Laine, A. (2026). Pictures in mathematical word problems: A systematic review on the role of task and student characteristics. *Educational Psychology Review*, 38(1), Article 20. <https://doi.org/10.1007/s10648-025-10112-5>
- Vessonen, T., Hellstrand, H., Kurkela, M., Aunio, P., & Laine, A. (2025). The effectiveness of mathematical word problem-solving interventions among elementary schoolers: A systematic review and meta-analysis. *International Journal of Educational Research*, 132, Article 102642. <https://doi.org/10.1016/j.ijer.2025.102642>
- Wilkie, K. J., & Hopkins, S. (2024). Generalizing actions with the subtraction-compensation property: Primary students’ algebraic thinking with tasks involving vertical towers of blocks. *Educational Studies in Mathematics*, 115(3), 433–458. <https://doi.org/10.1007/s10649-024-10303-x>

Xu, C., Di Lonardo Burr, S., Li, H., Liu, C., & Si, J. (2024). From whole numbers to fractions to word problems: Hierarchical relations in mathematics knowledge for Chinese Grade 6 students. *Journal of Experimental Child Psychology*, 242, Article 105884. <https://doi.org/10.1016/j.jecp.2024.105884>