ar-Riyadhiyat
JOURNAL OF MATHEMATICS EDUCATION
Vol. 7, No. 1, July 2026, pp. 31-39
ISSN: 2775-8990 DOI: 10.47766/arriyadhiyyat.v7i1.7581

Effect of PhET-Assisted Discovery Learning on Junior High School Students' Mathematical Conceptual Understanding

Janna Aprilia Azhara*, Muhammad Zaki, Guntur Maulana Muhammad
Department of Mathematics Education, Universitas Samudera, Langsa, Indonesia

Article Info

Article history:

Received April 18, 2026
Revised July 03, 2026
Accepted July 15, 2026
Published July 30, 2026

Keywords:

Conceptual understanding
Discovery learning
Mathematics education
PhET simulation
Statistics

ABSTRACT

Measures of central tendency are taught through procedures that may leave students with limited opportunities to explore their underlying relationships. This quasi-experimental study compared discovery learning supported by PhET simulations with discovery learning without simulation support in two intact Grade VIII classes at SMP Negeri 2 Langsa, Indonesia. The posttest-only design involved 50 students (25 per class), and conceptual understanding was assessed using essay responses about statistical concepts. The PhET-supported class had a higher mean posttest score ($M = 78.76$, $SD = 17.06$) than the comparison class ($M = 62.88$, $SD = 21.50$). The student-level independent-samples test yielded $t(48) = 2.89$, $p = .006$, a mean difference of 15.88 points (95% CI [4.84, 26.92]), and Cohen's $d = 0.82$. The result indicates higher observed performance in the PhET-supported class. However, with one intact class in each condition and no pretest, class-related differences cannot be separated from the instructional contrast. The findings offer a preliminary classroom-level rationale for examining simulation-supported discovery activities in statistics instruction using better-controlled designs.

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

Corresponding Author:

Janna Aprilia Azhara
e-mail: apriliajanna@gmail.com

INTRODUCTION

Measures of central tendency provide an important entry point into statistical reasoning because they require learners to relate individual observations to characteristics of an entire distribution. Understanding the mean, median, and mode goes beyond reproducing computational rules: students must interpret the context of data, connect numerical results with visual representations, and judge whether a summary is informative. Kurnia et al. (2024) emphasize that statistical literacy integrates representational and statistical-mathematical knowledge with interpretation and evaluation. Mathematics-education research likewise treats flexible movement between representations as integral to understanding, especially when digital and paper-based resources are combined (Weigand et al., 2024). These considerations make conceptual understanding, rather than formula recall alone, a relevant outcome for statistics instruction.

Conceptual understanding is also a foundation for the broader aims of mathematics education, which include developing logical, systematic, and analytical thinking and communicating ideas through symbols, diagrams, and mathematical language. A student who understands a concept can restate it in their own words, classify objects according to whether they satisfy the concept, present it in different representational forms, distinguish examples from non-examples, and select appropriate procedures for applying it. In statistics, these capabilities allow learners to explain why a particular measure summarizes a data set well, rather than simply computing a value. Instruction that emphasizes memorized steps may therefore leave students able to calculate a mean or locate a median but unable to interpret what the result reveals about the data or to judge when one measure is more appropriate than another.

In the classroom context of this study, preliminary observations and discussions with mathematics teachers at SMP Negeri 2 Langsa suggested that some Grade VIII students found measures of central tendency difficult and had limited opportunities to explore the content using interactive resources. The teachers also reported that student interest tended to decline during predominantly teacher-directed lessons and that visual or manipulable learning media were seldom used when the topic was taught. These observations were not a formal baseline assessment, but they motivated investigation of an instructional approach that allowed students to examine and explain patterns. Discovery learning was selected because it organizes inquiry through stimulation, problem identification, data collection, processing, verification, and generalization (Khasinah, 2021). Evidence syntheses have reported favorable mathematics outcomes for guided discovery approaches (Khairunnisa & Juandi, 2022; Khairunnisa, 2025), although broader reviews indicate that the quality of guidance matters more than unguided exploration itself (Alfieri et al., 2011; Lazonder & Harmsen, 2016).

Each stage of discovery learning serves a distinct instructional function. During stimulation and problem identification, students encounter a situation that arouses curiosity and prompts them to formulate questions. In data collection and data processing, they gather relevant information and look for patterns or relationships. Verification asks students to check whether the evidence supports their initial ideas, and generalization enables them to formulate broader conclusions that can be communicated to others. This sequence is compatible with concept-oriented mathematics learning because it engages students in observing, analyzing, checking, and explaining relationships rather than receiving them as finished statements. At the same time, the quality of what students can observe and verify depends partly on the resources available to them; when relationships are represented only in static examples, opportunities to test predictions across many cases may be limited.

Interactive simulations offer a complementary representational resource. Manipulating objects and observing changes can make relationships more visible, while a guided task can prompt learners to predict, test, and justify their interpretations. For measures of central tendency, such an environment could allow students to add, remove, or modify data values and immediately observe how the mean, median, and mode respond—for instance, noticing that an extreme value shifts the mean more than the median. Relationships of this kind can be difficult to convey through static explanations or a small number of worked examples alone. PhET was developed to support inquiry through manipulable models rather than passive demonstrations (Wieman et al., 2008). In mathematics settings, Arifin et al. (2023) studied PhET-supported work on mixed fractions, and Dwiyanti and Nurafni (2023) examined interactive simulation media for learning integer concepts. Findings from science classrooms also illustrate possible benefits of simulation-based exploration, but they cannot be transferred automatically to statistics lessons (Dy et al., 2024). Technology-enhanced mathematics learning is therefore best understood as an interaction among the task, the teacher's guidance, and learners' use of representations (Engelbrecht & Borba, 2024; Radović, 2024).

Reviews of digital resources in mathematics indicate that learning outcomes vary according to content, implementation, and assessment (Teapon et al., 2024; Rahma & Nurlaelah, 2024). For instance, other studies have described opportunities created by interactive mathematics resources (Aulia & Pramuditya, 2025) and manipulable representations (Masitoh & Prasetyawan, 2026). Yet these findings do not isolate the contribution of PhET to learning measures of central tendency within an otherwise shared discovery-learning model. Many prior studies also report final achievement scores without describing the facets of conceptual understanding that the assessment was intended to capture, such as classification, representation, example identification, and algorithmic application. Moreover, the educational value of simulation media is likely to depend on how they are embedded in a structured instructional model: access to a digital tool alone does not ensure that students will investigate, interpret, verify, and generalize the relationships they observe. A comparison in which both classes use the same learning approach, while one class receives PhET support, can provide relevant classroom-level evidence. This comparison remains observational with respect to differences between the two intact classes when only one class is observed in each condition.

Accordingly, we compared the posttest performance of students taught measures of central tendency through PhET-assisted discovery learning with the performance of students taught through discovery learning without PhET. We examined whether the observed difference in overall conceptual-understanding scores favored the PhET-supported class, and interpreted the findings in relation to the design's restrictions. The focus was the comparison of total posttest scores, not the causal identification of the digital tool's independent effect or a separate test of each conceptual indicator. Consistent with the instructional rationale, we expected the PhET-supported class to obtain higher total posttest scores than the comparison class, while recognizing that a single two-class comparison can provide only preliminary evidence on this expectation.

METHOD

Design and participants

We used a quantitative, quasi-experimental posttest-only comparison of two intact Grade VIII classes at SMP Negeri 2 Langsa during the second semester of the 2024–2025 academic year. This design was chosen because students were organized into existing classes, making individual random assignment impractical within the school's regular timetable. The Grade VIII population comprised 216 students across seven classes. Two classes were selected purposively using teacher information and prior classroom achievement as informal indicators of comparability; individual baseline test results and an experimental pretest were not reported. The two selected classes, each comprising 25 students, were assigned to instructional conditions at the class level: VIII-7 received PhET-assisted discovery learning and VIII-6 received discovery learning without PhET (see Table 1). We randomized the class-level condition designation after selecting the two existing classes; individual students were not randomized.

Table 1. Instructional conditions in the posttest-only comparison

Class / condition	Instructional approach	Posttest
VIII-7 / PhET supported (n = 25)	Discovery learning with PhET simulation	Same essay-based test
VIII-6 / Comparison (n = 25)	Discovery learning without PhET simulation	Same essay-based test

Both classes studied mean, median, and mode as part of the same curriculum. We planned comparable learning objectives, lesson time, and assessment procedures across conditions; the intended instructional difference was the use of simulations. Because each condition coincided with one intact class, any pre-existing differences between classes remain potentially confounded with the instructional condition. Lessons were delivered during regularly scheduled mathematics periods to preserve the natural classroom setting, and both classes continued to receive the content required by the school curriculum. The independent variable was therefore the instructional condition (discovery learning with or without PhET support), and the dependent variable was students' conceptual understanding of measures of central tendency, operationalized as their total posttest scores.

Learning procedures

Instruction in both classes followed the six discovery-learning stages: stimulation, problem identification, data collection, data processing, verification, and generalization (Khasinah, 2021). In both classes, the teacher acted as a facilitator who guided students through the investigative process rather than presenting definitions and formulas at the outset. In the PhET-supported class, learners used interactive simulation activities as part of examining data values and their relationships to measures of central tendency. These activities involved prediction, manipulation of displayed data, and comparison with hand calculations. During verification, students discussed whether computed measures and displayed patterns were consistent; they then formulated conclusions about the situations explored. In the comparison class, the corresponding stages relied on teacher-prepared problems, worksheets, discussion, and manual calculations without PhET.

Supporting materials comprised learning modules and student worksheets organized according to the discovery-learning stages. The worksheets for the PhET-supported class contained prompts directing students to explore measures of central tendency in the simulation, record what they observed, and compare the displayed values with their own calculations. The worksheets for the comparison class addressed the same learning objectives through teacher-prepared data sets, guiding questions, discussion activities, and manual calculations. Thus, both classes were expected to identify problems, collect and process information, verify results, and formulate conclusions; the intended difference concerned the resource used to explore the data rather than the sequence of learning activities (see Table 2).

The specific PhET simulation title, software version, screenshots, and lesson-by-lesson activity sequence are not provided in this article. Consequently, we describe the intended instructional comparison without assuming that one simulation contained every activity concerning mean, median, and mode. These implementation details are needed to replicate our lessons exactly.

Conceptual-understanding assessment

The outcome measure was an essay-based posttest addressing statistical measures of central tendency. Our assessment framework identified four aspects of conceptual understanding: classification of mathematical or data objects, representation of statistical information, recognition of relevant examples, and selection or application of appropriate algorithms. Students were expected to provide written solutions rather than select answers from a fixed-response format. We report overall test scores rather than an item-by-indicator blueprint

or indicator-level results; we therefore make no claim that the difference belongs to a particular conceptual dimension.

The four aspects were operationalized as follows. Classification referred to students' ability to group data or mathematical objects according to relevant statistical characteristics. Representation referred to organizing, presenting, and interpreting statistical information in tables, diagrams, or mathematical expressions. Example identification referred to recognizing situations or data sets that illustrate a given concept, as well as those that do not. Algorithmic application referred to selecting, carrying out, and explaining appropriate procedures for determining and interpreting the mean, median, and mode. Essay items were used because they allow students to display their reasoning and explain their procedures. Responses were scored with a predetermined rubric that considered the accuracy of the answer, the appropriateness of the representation or procedure, and the clarity of the reasoning provided.

Two expert validators—one focusing on mathematical content and one on learning media—reviewed the test and supporting materials before implementation. The content validator examined the accuracy of the mathematics, alignment with the curriculum and learning objectives, and correspondence between items and the intended aspects of conceptual understanding. The media validator examined the clarity of instructions, the suitability and visual presentation of the learning materials, and the appropriateness of the PhET-assisted activities for Grade VIII students. Both validators also commented on the clarity of language and the feasibility of classroom implementation, and the materials were revised in response to their comments before use.

Our instrument checks included Pearson item correlations, Cronbach's alpha, item difficulty, and item discrimination. The difficulty index was used to classify items as easy, moderate, or difficult, and the discrimination index to indicate how well each item distinguished students with higher and lower levels of understanding; items that did not meet the criteria were revised or excluded before the final posttest. We do not present the numerical item coefficients, overall alpha value, scoring weights, or complete test items in this article. Thus, we avoid numerical claims about measurement quality, and further documentation is needed for independent appraisal of the test.

Data collection and analysis

After the teaching activities, both classes completed the same posttest under comparable stated conditions. The same test instructions, completion time, scoring rubric, and administration procedures were applied in both classes, and completed scripts were scored using the predetermined rubric. The analysis summarized each class using its sample size, score range, mean, and standard deviation. Shapiro–Wilk tests, appropriate for groups of 25 students, assessed departures from normality within each class, and Levene's test assessed the variance assumption. The reported independent-samples t-test used an equal-variance model with 48 degrees of freedom and a two-sided significance threshold of .05. The mean difference, 95% confidence interval, and Cohen's d were calculated from the reported group summaries, with d defined as the mean difference divided by the pooled standard deviation.

The inferential test treats the 50 student scores as independent observations. Its p value is conditional on the assumption that student observations are independent. With only one intact class per condition, class effects cannot be separated from the intervention and cluster-level uncertainty cannot be estimated reliably. Therefore, the test statistics are presented as descriptive inferential summaries of the observed student-level contrast, not as definitive causal evidence for PhET.

Research ethics

Permission for classroom data collection was obtained from SMP Negeri 2 Langsa, and individual student identities were not disclosed in the reported results. Before data collection, the school administration and the mathematics teacher were informed about the study objectives, instructional procedures, and data-collection process. The posttest data were used solely for research purposes, and both classes received the curriculum content required by the school as part of regular lessons. The available study documentation does not describe parental or guardian consent procedures or a formal ethics-review determination; we therefore do not make claims about safeguards beyond the school permission stated above.

Table 2. Classroom instruction reported for the two conditions

Discovery phase	PhET-supported classroom	Classroom without PhET
Stimulation	Observe data-related situations in the simulation	Examine teacher-prepared statistical examples
Problem identification	Pose questions about mean, median, and mode	Pose questions using worksheets and examples
Data collection	Manipulate values and record displayed outcomes	Gather values and examples through worksheets

Discovery phase	PhET-supported classroom	Classroom without PhET
Data processing	Identify patterns in the recorded data	Compute statistics manually and discuss patterns
Verification	Compare displayed outcomes with hand calculations	Check calculations and proposed explanations
Generalization	Summarize relationships observed across cases	Draw conclusions from non-digital cases

RESULTS AND DISCUSSION

Group-level posttest performance

Both classes comprised 25 students. The PhET-supported class scored higher on the posttest (M = 78.76, SD = 17.06) than the discovery-learning comparison class (M = 62.88, SD = 21.50). The observed difference was 15.88 points. The two distributions overlapped, as reflected in the respective ranges of 44–100 and 28–97 (see Table 3).

Table 3. Descriptive statistics for total posttest scores

Classroom condition	Students (n)	Minimum	Maximum	Mean	SD
PhET-supported discovery learning	25	44	100	78.76	17.06
Discovery learning without PhET	25	28	97	62.88	21.50

Beyond the difference in means, the descriptive statistics show that the PhET-supported class had a higher minimum score (44 vs. 28) and a smaller standard deviation, indicating somewhat less dispersion in its scores. Both classes nevertheless included students who scored near the maximum, and the ranges overlapped substantially. These descriptive patterns indicate a higher overall level of performance in the PhET-supported class, but inferential analysis was needed to judge whether the observed difference was larger than would be expected from within-class variation under the stated model.

Assumption checks and group comparison

The Shapiro–Wilk tests did not reject normality in either class (p = .076 and p = .061). Levene's test likewise did not reject the equal-variance assumption (p = .142). These nonsignificant results do not establish normality or equality of variances but were consistent with the assumptions used in the reported student-level comparison (see Table 4).

Both Shapiro–Wilk p values were relatively close to the .05 threshold, so the normality checks should be read cautiously. With 25 observations in each class, the independent-samples t-test is generally robust to moderate departures from normality. In addition, the comparison did not depend on the equal-variance assumption: a Welch test computed from the same summaries produced a nearly identical result, $t(45.64)$ = 2.89, p = .006.

Table 4. Assumption-test results

Test / group	Statistic	df	*p*
Shapiro–Wilk / PhET-supported	W = .928	25	.076
Shapiro–Wilk / comparison	W = .923	25	.061
Levene / between groups	F = 2.233	1, 48	.142

The equal-variance independent-samples analysis yielded $t(48)$ = 2.892, p = .006, with an estimated mean difference of 15.88 points and a 95% confidence interval of approximately [4.84, 26.92]. Cohen's d was 0.82, conventionally characterized as a large standardized difference. The mean difference, test statistic, confidence interval, and effect size were internally consistent when recalculated from the means, standard deviations, and sample sizes (see Table 5).

Table 5. Student-level comparison of total posttest scores

Comparison	t (df)	*p*	Mean difference	95% CI	Cohen's *d*
PhET-supported − comparison	2.892 (48)	.006	15.88	[4.84, 26.92]	0.82

The pooled standard deviation was approximately 19.41, so the observed mean difference corresponded to about four-fifths of a pooled standard deviation. Because the 95% confidence interval excluded zero, the student-level analysis did not support a zero mean difference. The interval was nevertheless wide,

ranging from a modest difference of about 5 points to a substantial difference of about 27 points, which reflects the limited sample size and the considerable within-class variation.

These results describe a favorable observed score difference for the PhET-supported class. They do not identify the relative performance of the four conceptual indicators, because corresponding subscale summaries were not reported, nor do they exclude baseline class differences as an alternative explanation.

Discussion

The principal empirical finding is a higher average score in the class receiving discovery learning with PhET simulation. The magnitude of the contrast was substantial relative to the within-class score variation. Because both classes were described as following discovery-learning stages, the comparison is more specific than a contrast between any digital instruction and an entirely different conventional method. Nevertheless, it is a comparison of two intact classes, not a replicated trial of PhET across multiple classes. The positive student-level difference is therefore consistent with a possible instructional benefit but does not establish that the simulation caused the difference.

The finding can be read in relation to constructivist views of learning, in which conceptual understanding develops as learners actively construct knowledge and connect new information with what they already know. Discovery learning operationalizes this view through stages that ask students to explore, identify problems, collect and process data, verify their ideas, and generalize. Within that shared framework, the PhET-supported class had an additional means of generating and examining cases: students could alter values and see the consequences for the mean, median, and mode rather than relying mainly on symbolic or verbal explanations. Whether this additional resource accounts for the score difference cannot be determined from the present design, but the interpretation is consistent with the instructional rationale for the comparison.

One plausible educational interpretation concerns the relationship between exploratory activities and representations of data. Simulations allow some mathematical quantities to be varied while outcomes are displayed, potentially supporting comparison, prediction, and explanation. PhET's inquiry-oriented design provides a rationale for this type of activity (Wieman et al., 2008), while studies of mathematics instruction have documented the learning opportunities associated with interactively represented fractions and integers (Arifin et al., 2023; Dwiyanti & Nurafni, 2023). The evidence in those studies differs in grade level and content, however. In the present investigation, students' actual interactions with the simulation were not recorded as process data, so this representational pathway remains a teaching rationale rather than a measured mechanism.

A complementary interpretation concerns the experiences available in the comparison class. Students there followed the same stages but worked with teacher-prepared examples, worksheets, and board explanations. Such static representations can support careful calculation and discussion, yet they offer fewer opportunities to generate many cases quickly and to see how a change in one value propagates through each measure. Visual and interactive support might also reduce the effort students spend on repeated computation, allowing more attention to relationships among measures. These possibilities are plausible but untested here, because we did not measure cognitive load, time on task, or the number of cases students examined in either class.

The discovery-learning framework is similarly important. Its stages provide an opportunity for students to interpret tasks, check predictions, and communicate conclusions, but positive outcomes depend on appropriate guidance. Alfieri et al. (2011) distinguished assisted discovery from unguided discovery, and Lazonder and Harmsen (2016) identified facilitative effects of guidance in inquiry-based learning. Mathematics-focused syntheses report that discovery learning can support mathematical competencies (Khairunnisa & Juandi, 2022; Khairunnisa, 2025), while related guided-inquiry research has also reported favorable outcomes (Nurtamam et al., 2024). These results should not be interpreted as explaining the present score difference in isolation: discovery learning was provided to both classes, so the comparison concerns the added instructional resource and any other class-specific differences accompanying its use. A meta-analysis of GeoGebra-assisted discovery learning further illustrates the potential of digital resources when they are incorporated into a defined learning model, although it concerns a different technology and mathematical outcome (Anjarwati et al., 2022).

Recent technology-focused mathematics research similarly emphasizes conditions of implementation rather than treating digital tools as universally beneficial. Studies have reported favorable perceptions of interactive media for specific mathematics topics (Aulia & Pramuditya, 2025), and research on conceptual understanding in another teaching model points to the value of contextualized conceptual work (Putri et al., 2025). Reviews of mathematics technology demonstrate diversity in tools and outcomes (Engelbrecht & Borba, 2024; Teapon et al., 2024; Weigand et al., 2024). Meta-analytic evidence from Indonesian secondary mathematics also suggests that effects differ among topics and levels (Rahma & Nurlaelah, 2024). Accordingly,

the statistical topic, exact PhET application, form of teacher scaffolding, and assessment tasks should be specified before general recommendations about simulation-assisted discovery learning are made.

The teacher's role deserves particular attention in this respect. A simulation can display many cases quickly, but students may not notice the intended relationships unless tasks direct their attention, ask them to commit to predictions, and require them to explain discrepancies between what they expected and what they observed. In the present lessons, the teacher acted as a facilitator in both classes, and the worksheets were organized around the same discovery stages. Future implementations could make this guidance more explicit by documenting the prompts used at each stage, the questions teachers asked during verification, and the ways in which students' conclusions were discussed and refined during generalization. Such documentation would help distinguish the contribution of the simulation from the contribution of the guidance that surrounded its use.

For assessment, the total-score result needs to be distinguished from the underlying construct. Conceptual understanding of mean, median, and mode includes explaining why a measure fits a distribution and interpreting representations, not simply choosing or executing algorithms. Kurnia et al. (2024) describe multiple components of statistical literacy, reinforcing the need for tasks that assess interpretation as well as calculation. We did not present examples or separate scores for each of the four conceptual indicators. A future application should report the item blueprint, scoring anchors, reviewer evidence, item correlations, reliability estimate, and example responses so readers can judge whether improved total scores represent deeper understanding across concepts rather than stronger performance on a subset of procedures.

Such reporting would also make it possible to examine common difficulties with measures of central tendency that a total score can conceal. Students may, for example, locate the median as the middle value of an unordered list, treat the mean as the most appropriate summary regardless of the shape of the distribution, or overlook how a single extreme value affects the mean but not the median. Items that target these difficulties, together with students' written explanations, would indicate whether simulation-supported exploration helps learners revise such ideas or mainly improves the accuracy of routine calculations. This distinction is central to judging whether gains reflect conceptual understanding in the sense described in the Introduction.

Several design limits materially constrain interpretation. First, the absence of a pretest means that baseline comparability was based on teacher impressions and unspecified prior achievement records. Second, assigning a single intact class to each condition creates a complete overlap between class and instruction; any difference in students, peer interactions, classroom conditions, or teacher delivery may contribute to the contrast. The usual independent-samples t-test does not correct that class-level confounding. Third, we have not fully described the specific PhET application, the intervention sequence, and the assessment's psychometric results. Fourth, we did not report engagement, interaction logs, or qualitative reasoning evidence, limiting what can be inferred about why the classes differed. Multiclass designs with a documented baseline, cluster-aware analysis, and instrument-level reporting would provide stronger evidence.

For classroom practice, a prudent implication is to trial carefully guided simulation activities as one resource for exploring central tendency, while also using manual calculations, discussion, and explicit interpretation tasks. Teachers could ask students to predict the effect of changing observations, compare measures that respond differently to extreme values, and justify their preferred summary for a given distribution. These are suggestions for future instructional design rather than a claim that every example was implemented in the study lessons. Further research can test the value of those features with a sufficiently replicated design rather than attributing a two-class difference exclusively to the presence of PhET.

Such research would benefit from larger samples drawn from multiple schools, pretest–posttest or other baseline measures that allow learning gains to be estimated, and additional outcomes such as student engagement, motivation, and the quality of written statistical reasoning. Extending the comparison to other statistical and mathematical topics would also clarify whether any benefit is specific to measures of central tendency or extends to content in which dynamic representations play a similar role.

CONCLUSION

In this study, the Grade VIII class receiving PhET-assisted discovery learning obtained higher total posttest scores on measures of central tendency than the class receiving discovery learning without simulation support. The difference was sizable in our student-level analysis, but the posttest-only comparison of one intact class per condition does not permit a firm causal conclusion. The results offer preliminary, context-specific support for examining interactive simulations as part of guided statistical exploration. More rigorous multiclass research, baseline assessment, instrument documentation, and records of how students use the digital environment are needed to determine whether and through what processes the simulation contributes to conceptual understanding.

From an instructional-design perspective, the study illustrates one way of combining a student-centered learning model with an interactive digital resource while keeping the learning sequence constant across classes. Teachers who adopt this approach should treat the simulation as a tool for generating and checking conjectures, not as a substitute for explanation, calculation, or discussion. Future studies should involve larger and more diverse samples across schools, include baseline measures, report indicator-level and item-level evidence, and examine engagement and reasoning processes so that the conditions under which simulation-assisted discovery learning supports conceptual understanding can be specified more precisely.

ACKNOWLEDGMENTS

We thank the participating school, mathematics teachers, and students at SMP Negeri 2 Langsa for supporting classroom implementation and the collection of learning-assessment data. We also acknowledge Nursamsu, S.Pd., M.Pd., for his administrative support.

DECLARATIONS

Author Contributions: Janna Aprilia Azhara: Conceptualization, Methodology, Investigation, Data Curation, Formal Analysis, Writing – Original Draft. Muhammad Zaki: Conceptualization, Methodology, Validation, Supervision, Writing – Review & Editing. Guntur Maulana Muhammad: Methodology, Formal Analysis, Supervision. All authors have read and approved the final manuscript.
Funding: This research received no external funding.
Conflict of Interest: The authors declare no conflict of interest.
Ethical Approval and Informed Consent: Permission to conduct the study and collect classroom data was obtained from SMP Negeri 2 Langsa. Student identities were kept confidential and are not disclosed in the reported results. The study was not reviewed by a formal research ethics committee.
Use of Artificial Intelligence: During the preparation of this manuscript, the authors used ChatGPT and Grammarly to improve the language and English writing. The authors reviewed and edited the output and take full responsibility for the content of the publication.

REFERENCES

Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology*, *103*(1), 1–18. https://doi.org/10.1037/a0021017

Anjarwati, D., Juandi, D., Nurlaelah, E., & Hasanah, A. (2022). Studi meta-analisis: Pengaruh model discovery learning berbantuan GeoGebra terhadap kemampuan berpikir kritis matematis siswa. *Jurnal Cendekia: Jurnal Pendidikan Matematika*, *6*(3), 2417–2427. https://doi.org/10.31004/cendekia.v6i3.1506

Arifin, S., Razali, F. B., & Rahayu, W. (2023). Integrating PhET interactive simulations to enhance students' mathematical understanding and engagement in learning mixed fractions. *Al Ibtida: Jurnal Pendidikan Guru MI*, *10*(2), 241–252. https://doi.org/10.24235/al.ibtida.snj.v10i2.15056

Aulia, G. P., & Pramuditya, S. A. (2025). Interactive learning media using Google Sites to enhance students' understanding of ratio concepts in Grade VII. *Ar-Riyadhiyyat: Journal of Mathematics Education*, *6*(1), 30–38. https://doi.org/10.47766/arriyadhiyyat.v6i1.5923

Dwiyanti, I. R., & Nurafni. (2023). PhET interactive simulation media improves students' understanding of mathematical concepts in integer material. *Mimbar PGSD Undiksha*, *11*(2), 191–197. https://doi.org/10.23887/jjpgsd.v11i2.64417

Dy, A. U., Lagura, J. C., & Baluyos, G. R. (2024). Using PhET interactive simulations to improve the learners' performance in science. *EduLine: Journal of Education and Learning Innovation*, *4*(4), 520–530. https://doi.org/10.35877/454RI.eduline2981

Engelbrecht, J., & Borba, M. C. (2024). Recent developments in using digital technology in mathematics education. *ZDM – Mathematics Education*, *56*, 281–292. https://doi.org/10.1007/s11858-023-01530-2

Khairunnisa, K. (2025). A meta-analysis of the effect of discovery learning model on students' conceptual understanding and mathematical reasoning abilities. *International Journal of Education*, *18*(1). https://ejournal.upi.edu/index.php/ije/article/view/54959/0

Khairunnisa, K., & Juandi, D. (2022). Meta-analysis: The effect of discovery learning models on students' mathematical ability. *Jurnal Riset Pendidikan Matematika*, *9*(2), 201–211. https://doi.org/10.21831/jrpm.v9i2.49147

Khasinah, S. (2021). Discovery learning: Definisi, sintaksis, keunggulan dan kelemahan. *Jurnal Mudarrisuna: Media Kajian Pendidikan Agama Islam*, *11*(3), 402–413. https://doi.org/10.22373/jm.v11i3.5821

Kurnia, A. B., Lowrie, T., & Patahuddin, S. M. (2024). The development of high school students' statistical literacy across grade level. *Mathematics Education Research Journal*, *36*(Suppl. 1), 7–35. https://doi.org/10.1007/s13394-023-00449-x

Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning: Effects of guidance. *Review of Educational Research*, *86*(3), 681–718. https://doi.org/10.3102/0034654315627366

Masitoh, L. F., & Prasetyawan, E. (2026). Enhancing mathematics achievement through virtual manipulatives: A meta-analysis of K–9 intervention. *AL-ISHLAH: Jurnal Pendidikan*, *17*(4), 7034–7047. https://doi.org/10.35445/alishlah.v17i4.7586

Nurtamam, M. E., Budayasa, I. K., & Lukito, A. (2024). Process-oriented guided inquiry learning and students' mathematical reasoning ability in Indonesia: A systematic literature review and meta-analysis. *Tadris: Jurnal Keguruan dan Ilmu Tarbiyah*, *9*(2), 347–358. https://doi.org/10.24042/tadris.v9i2.21308

Putri, A., Yunita, A., & Mardiyah, A. (2025). Enhancing students' conceptual understanding in mathematics through the realistic mathematics education model. *Ar-Riyadhiyyat: Journal of Mathematics Education*, *6*(1), 47–53. https://doi.org/10.47766/arriyadhiyyat.v6i1.5964

Radović, S. (2024). Is it only about technology? The interplay between educational technology for mathematics homework, teaching practice, and students' activities. *Journal of Computers in Education*, *11*, 743–762. https://doi.org/10.1007/s40692-023-00277-9

Rahma, F. A., & Nurlaelah, E. (2024). Effectiveness of digital technology on enhancing mathematics achievement of Indonesian secondary schools students: A meta-analysis research from 2018–2023. *Jurnal Pendidikan MIPA*, *25*(2), 803–813. https://doi.org/10.23960/jpmipa/v25i2.pp803-813

Teapon, N., Kusumah, Y. S., Pakaya, Y. M., Adjis, V. A., & Sartika, R. A. (2024). Heterogeneity of interactive learning media outcomes for improving mathematical competence: A systematic literature review. *Mathline: Jurnal Matematika dan Pendidikan Matematika*, *9*(1), 205–216. https://doi.org/10.31943/mathline.v9i1.575

Weigand, H.-G., Trgalová, J., & Tabach, M. (2024). Mathematics teaching, learning, and assessment in the digital age. *ZDM – Mathematics Education*, *56*, 525–541. https://doi.org/10.1007/s11858-024-01612-9

Wieman, C. E., Adams, W. K., & Perkins, K. K. (2008). PhET: Simulations that enhance learning. *Science*, *322*(5902), 682–683. https://doi.org/10.1126/science.1161948