JARME Journal of Authentic Research on Mathematics Education
Volume 8, No. 2, July 2026
https://doi.org/ 10.37058/jarme.v3i1.18467 e-ISSN 2655-7762

E-Worksheet-Assisted Problem-Based Learning and Students' Mathematical Critical Thinking in Measures of Central Tendency

Riana Putri Nathania [1], Widya Dwiyanti [1*], Ucu Koswara [1]

[1]Universitas Sebelas April, Jl. Angkrek Situ No.19 Kec. Sumedang Utara, Kabupaten Sumedang, Jawa Barat, 45323, Indonesia.
E-mail: widdwiyanti@unsap.ac.id*

ARTICLE INFO

Article history
Received: 2026-01-04
Revised: 2026-03-05
Accepted: 2026-08-04

Keywords
Berpikir kritis matematis, e-LKPD, pembelajaran berbasis masalah, ukuran pemusatan data

E-worksheet, mathematical critical thinking, measures of central tendency, problem-based learning

ABSTRACT

Pada ukuran pemusatan data, siswa perlu menilai karakteristik data dan memberikan alasan atas pilihan ukuran yang digunakan. Proses ini berkaitan dengan kemampuan berpikir kritis matematis. Penelitian ini mengkaji peningkatan kemampuan berpikir kritis matematis melalui problem-based learning (PBL) berbantuan E-LKPD serta pola capaian indikator. Penelitian kuasi-eksperimen ini melibatkan 40 siswa dari dua kelas nonekuivalen (16 eksperimen; 24 kontrol). Kelas dipilih secara purposif berdasarkan rata-rata nilai matematika semester sebelumnya yang relatif serupa. Kelompok eksperimen memperoleh PBL berbantuan E-LKPD, sedangkan kontrol memperoleh pembelajaran ekspositori konvensional. Data tes uraian empat butir dianalisis menggunakan N-gain, uji t dua sampel independen, Hedges' g, dan ANCOVA tambahan. Rata-rata N-gain kelompok eksperimen lebih tinggi (0,548 vs. 0,416), $t(38) = 2{,}731$, $p = 0{,}010$, dengan Hedges' $g = 0{,}864$ yang secara konvensional dikategorikan besar. Namun, perbedaan posttest setelah memperhitungkan skor awal tidak signifikan, $F(1, 37) = 2{,}664$, $p = 0{,}111$. Secara deskriptif, kelompok eksperimen lebih tinggi pada evaluasi dan inferensi, kontrol pada interpretasi, sedangkan analisis relatif serupa. Berdasarkan N-gain, kelompok eksperimen menunjukkan peningkatan kemampuan berpikir kritis matematis yang lebih tinggi. Namun, bukti belum cukup untuk menyimpulkan keunggulan setelah penyesuaian skor awal. Temuan terbatas pada siswa kelas VIII dan materi ukuran pemusatan data.

In measures of central tendency, choosing an appropriate measure requires evaluating data characteristics and justifying the choice. These demands involve mathematical critical thinking. This study examined improvement in mathematical critical thinking under e-worksheet-assisted problem-based learning (PBL) and described performance across indicators. A quasi-experimental nonequivalent control group design included 40 students from two purposively selected classes (16 experimental; 24 control). The classes had similar previous-semester mathematics report-card averages. The experimental group received e-worksheet-assisted PBL; control students received conventional expository instruction. Data were *collected using a four-item constructed-response test. Analyses included N-gain, an independent-samples t test, Hedges' g, and ANCOVA. The experimental group showed a higher mean N-gain (0.548 vs. 0.416), t(38) = 2.731, p = .010. Hedges' g was 0.864, conventionally classified as large. However, the baseline-adjusted posttest difference was not significant, F(1, 37) = 2.664, p = .111. Descriptively, the experimental group scored higher on evaluation and inference. The control group scored higher on interpretation, while analysis scores were similar. Based on N-gain, the experimental group showed greater improvement in mathematical critical thinking. However, the evidence was insufficient to conclude superiority after baseline adjustment. These findings are limited to eighth-grade students and measures of central tendency.*

How to Cite:

Nathania, R. P., Dwiyanti, W., & Koswara, U. (2026). E-Worksheet-Assisted Problem-Based Learning and Students' Mathematical Critical Thinking in Measures of Central Tendency. ***Journal of Authentic Research on Mathematics Education, 5***(1), 1-3. https://doi.org/10.37058/jarme.v3i1.18467

1. INTRODUCTION

Critical thinking is widely recognized as an important learning goal across subject areas. Within mathematics, it involves interpreting information, analyzing relationships, evaluating claims and solution processes, and explaining or justifying conclusions (Monteleone et al., 2023). Facione (2015) identified six core critical thinking skills: interpretation, analysis, evaluation, inference, explanation, and self-regulation. Drawing on this framework, the present study focused on four of these skills: interpretation, analysis, evaluation, and inference.

These skills are particularly important in statistics, where numerical results must be interpreted and evaluated before they can support a conclusion. From a critical thinking perspective, statistical literacy involves identifying statistical claims and evidence, assessing the credibility of that evidence, determining whether the evidence supports those claims, considering alternative interpretations, and drawing conclusions from data (Koga, 2022). In the context of measures of central tendency, students engage in these processes when they interpret the mean, median, and mode, examine the characteristics of a data set, and decide which measure best represents the data or provides the most appropriate basis for a conclusion.

However, translating these critical thinking skills into classroom practice remains a challenge. Preliminary classroom observations at a public junior high school in Sumedang Regency indicated a contrast between students' procedural performance and their ability to reason about measures of central tendency in context. Students were generally able to calculate the mean, median, and mode, but had difficulty determining which measure was most appropriate for a given data set. In tasks involving extreme values, students tended to select the mean without adequately considering how those values affected its representativeness. Similar difficulties have been reported in previous studies. Karaca and Ay (2025) found that eighth-grade students demonstrated procedural knowledge of measures of central tendency but struggled to explain their meaning, interpret changes in data, draw inferences, and identify an appropriate measure to represent a data set. At the university level, Rosidah and Ikram (2021) similarly found that students often provided insufficient justification for their choice of measure and did not adequately consider the influence of outliers or the limitations of the mean, median, and mode. Taken together, the preliminary observations and previous findings highlight a gap between calculating measures of central tendency and using them to interpret, evaluate, and draw reasoned conclusions from data.

In problem-based learning (PBL), problems provide the organizing context for learning. With facilitation, students clarify the problem, recognize what they still need to learn, develop and evaluate possible solutions, and reflect on what they have learned (Hmelo-Silver, 2004). These activities create opportunities for mathematical critical thinking, although the quality of that engagement depends partly on the guidance provided. Evidence from inquiry-oriented learning likewise indicates that guidance can make student investigation more productive and can also be delivered through digital scaffolds (de Jong et al., 2023). In this study, the e-worksheet was designed to provide such support within the PBL sequence. Contextual problems and data representations were intended to focus students on relevant information. Guiding questions, activity instructions, and conclusion prompts were designed to structure solution development and make students' thinking more explicit. For measures of central tendency, these features were aligned with the need to interpret data, examine relationships or changes within a data set, evaluate whether a statistical measure is appropriate and whether the accompanying justification is adequate, and draw conclusions from the data. Related evidence comes from Muslimin et al. (2024), who embedded procedural prompts for investigation and hints for data analysis within PBL through e-scaffolding. After controlling for prior performance, students receiving the e-scaffolding showed higher HOTS scores than those receiving PBL without it. In the present study, the e-worksheet was positioned as pedagogical guidance embedded within PBL rather than simply as a digital format for presenting learning materials.

Previous research points to two related developments in mathematics learning. Digital worksheets have been designed to organize statistical activities intended to engage students in critical thinking. Febriani et al. (2024), for example, developed a flip-worksheet on the mean, median, and mode that combined infographics, activities, and open-ended questions. Digital resources have also been incorporated into PBL. Safaatullah et al. (2025), for instance, used Google Sites to support contextual problems, learning materials, videos, and practice activities in mathematics instruction. Closer examples can be found in higher education. Muerza et al. (2024) integrated student-created videos into PBL for probability and statistical inference. Tempelaar et al. (2024) combined PBL in an introductory mathematics and statistics course with online tutoring systems that provided problems, hints, worked solutions, and feedback.

The evidence does not yet converge on the instructional configuration examined here. Febriani et al. studied a digital worksheet for measures of central tendency through a development design rather than a comparison of instructional conditions. Safaatullah et al. examined digitally supported PBL and mathematical critical thinking, but on the Pythagorean theorem and with a posttest-only design. Muerza et al. (2024) and Tempelaar et al. (2024) provide examples of digital resources used alongside PBL in statistics-related higher education settings, but neither examined e-worksheet guidance for mathematical critical thinking at the lower-secondary level. More broadly, Ayari et al. (2025) found that digital-based approaches were prominent among integrated forms of PBL in K–12 mathematics, while data, probability, and statistics were among the less frequently investigated domains. Taken together, the literature supports the use of digital resources within PBL and in statistics learning. However, comparative evidence remains limited for an e-worksheet functioning as structured guidance within PBL on measures of central tendency. What also remains unclear is whether overall differences in mathematical critical thinking are accompanied by similar or contrasting patterns across the specific indicators assessed.

The present study addresses these issues by examining improvement in mathematical critical thinking under e-worksheet-assisted PBL, with conventional expository instruction serving as the comparison condition. It also describes performance patterns across the four indicators assessed. Conceptually, the study links the design of the e-worksheet to its role as structured guidance within PBL and to the thinking demands involved in working with measures of central tendency. Empirically, it adds comparative evidence on e-worksheet-assisted PBL in a statistical topic that remains relatively underrepresented in K–12 PBL research. To systematically address these objectives, the present study is guided by the following research questions:

1. Which group shows greater improvement in mathematical critical thinking: students receiving e-worksheet-assisted PBL or those receiving conventional expository instruction?
2. What descriptive patterns of mathematical critical thinking performance are observed across the assessed indicators in the two groups?

2. METHOD

2.1. Research Design and Participants

This study employed a quasi-experimental nonequivalent control group design. The experimental group received e-worksheet-assisted problem-based learning (PBL), whereas the control group received conventional expository instruction. Because intact classes were used without random assignment, pre-existing differences between the groups remained a potential threat to internal validity.

The study was conducted during the second semester of the 2024/2025 academic year at a public junior high school in Sumedang Regency, West Java, Indonesia. The population comprised 61 eighth-grade students in three intact classes. Two classes were purposively selected based on their similar first-semester mathematics report-card averages (81.35 for Class VIII-A and 81.21 for Class VIII-B) and their previous instruction by the same mathematics teacher. Class VIII-B (n = 16) was assigned as the experimental group and Class VIII-A (n = 24) as the control group. A pretest was administered to characterize baseline mathematical critical thinking performance. The same researcher taught both groups. Mathematical content, number of instructional sessions, and instructional time were kept consistent to reduce variation unrelated to the instructional conditions. Permission was obtained from the school before the study.

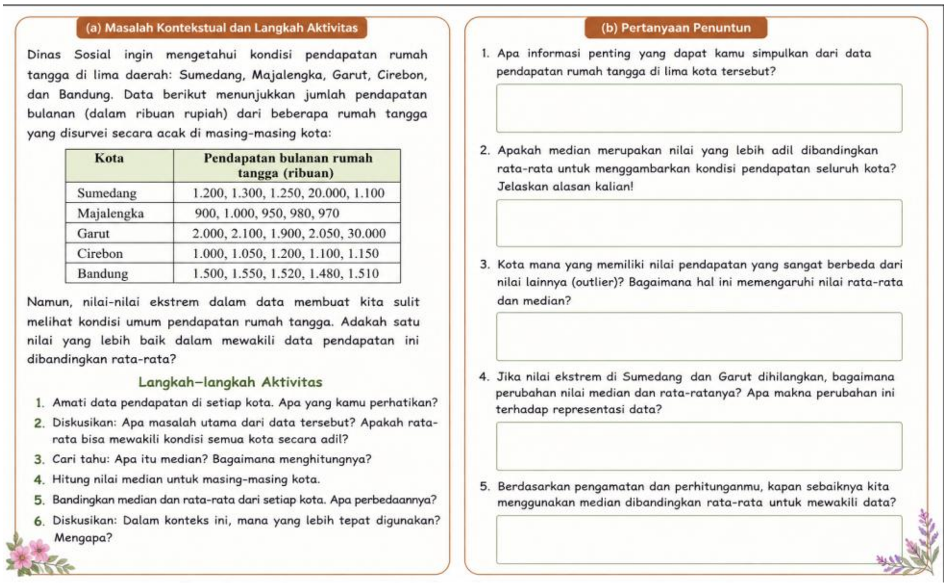
2.2. Research Procedures

The study consisted of five sessions in each group: a pretest, three instructional sessions on the arithmetic mean, median, and mode, and a posttest. Each instructional session lasted 80 minutes (2 × 40-minute periods), while the pretest and posttest each lasted 60 minutes.

Experimental-group instruction followed the five PBL stages proposed by Arends (2012): orienting students to the problem, organizing students for learning, guiding group investigation, developing and presenting their work, and analyzing and evaluating the problem-solving process. In each session, students worked collaboratively on a contextual problem, identified relevant information and task requirements, investigated and discussed possible solutions, presented their work, and reviewed the solution process and conclusions.

The PBL activities were supported by e-worksheets developed on the Liveworksheets platform and integrated into the PBL stages. Each e-worksheet contained an instructional video, a contextual problem, data representations, activity instructions, guiding questions, spaces for recording solutions, and a conclusion section. Figure 1 presents an example of a contextual problem and guiding questions from the e-worksheet on the median.

The control group received conventional expository instruction through whole-class explanation, worked examples, individual textbook exercises, and class discussion. Prepared instructional sequences and condition-specific materials were used to support procedural consistency across sessions. No independent observer or fidelity checklist was used; therefore, treatment fidelity was not quantified.



(a) Masalah Kontekstual dan Langkah Aktivitas

Dinas Sosial ingin mengetahui kondisi pendapatan rumah tangga di lima daerah: Sumedang, Majalengka, Garut, Cirebon, dan Bandung. Data berikut menunjukkan jumlah pendapatan bulanan (dalam ribuan rupiah) dari beberapa rumah tangga yang disurvei secara acak di masing-masing kota:

Kota	Pendapatan bulanan rumah tangga (ribuan)
Sumedang	1.200, 1.300, 1.250, 20.000, 1.100
Majalengka	900, 1.000, 950, 980, 970
Garut	2.000, 2.100, 1.900, 2.050, 30.000
Cirebon	1.000, 1.050, 1.200, 1.100, 1.150
Bandung	1.500, 1.550, 1.520, 1.480, 1.510

Namun, nilai-nilai ekstrem dalam data membuat kita sulit melihat kondisi umum pendapatan rumah tangga. Adakah satu nilai yang lebih baik dalam mewakili data pendapatan ini dibandingkan rata-rata?

Langkah–langkah Aktivitas

1. Amati data pendapatan di setiap kota. Apa yang kamu perhatikan?
2. Diskusikan: Apa masalah utama dari data tersebut? Apakah rata-rata bisa mewakili kondisi semua kota secara adil?
3. Cari tahu: Apa itu median? Bagaimana menghitungnya?
4. Hitung nilai median untuk masing-masing kota.
5. Bandingkan median dan rata-rata dari setiap kota. Apa perbedaannya?
6. Diskusikan: Dalam konteks ini, mana yang lebih tepat digunakan? Mengapa?

(b) Pertanyaan Penuntun

1. Apa informasi penting yang dapat kamu simpulkan dari data pendapatan rumah tangga di lima kota tersebut?
2. Apakah median merupakan nilai yang lebih adil dibandingkan rata-rata untuk menggambarkan kondisi pendapatan seluruh kota? Jelaskan alasan kalian!
3. Kota mana yang memiliki nilai pendapatan yang sangat berbeda dari nilai lainnya (outlier)? Bagaimana hal ini memengaruhi nilai rata-rata dan median?
4. Jika nilai ekstrem di Sumedang dan Garut dihilangkan, bagaimana perubahan nilai median dan rata-ratanya? Apa makna perubahan ini terhadap representasi data?
5. Berdasarkan pengamatan dan perhitunganmu, kapan sebaiknya kita menggunakan median dibandingkan rata-rata untuk mewakili data?

Figure 1. Example of an E-Worksheet Activity on the Median

2.3. Instrument and Data Collection

Data were collected using a constructed-response test of mathematical critical thinking on measures of central tendency. The test comprised four items, each designed to primarily elicit one of four critical thinking skills drawn from Facione's (2015) framework: interpretation, analysis, evaluation, and inference. Item scores were used descriptively to characterize students' performance on the corresponding indicators.

The same instrument and scoring guide were used for the pretest and posttest. Students' responses were evaluated using an analytic scoring rubric aligned with the indicator and the specific demands of each item. Each item had a maximum score of 16, resulting in a maximum total score of 64. The test specifications are presented in Table 1.

Table 1. Specifications of the Mathematical Critical Thinking Test

Indicator	Primary Task Demand
Interpretation	Interpreting information or the meaning of a measure of central tendency presented in the problem
Analysis	Examining relationships, conditions, or changes within a data set
Evaluation	Assessing the appropriateness of a choice, statement, or justification based on the data
Inference	Drawing a conclusion supported by an analysis of the data

The instrument and scoring guide were reviewed by two mathematics education experts using an 11-criterion validation form covering content, construction, and language. Each criterion was rated on a four-point scale. The expert ratings were quantified using Aiken's V, which ranged from 0.667 to 1.000 across the criteria, with a mean coefficient of 0.864. Both experts judged the instrument usable with minor revisions. Their feedback addressed the clarity of the scoring rubric, wording, and graphical information, and the instrument was revised accordingly before data collection. No empirical reliability coefficient was estimated for the test in the present study.

2.4. Data Analysis

The analysis began with descriptive statistics for the pretest and posttest scores in both groups. The reported statistics included sample size, mean, standard deviation, and minimum and maximum scores. Individual item scores were also analyzed descriptively to characterize performance patterns across the assessed indicators. The analysis included the mean pretest score, mean posttest score, mean score increase, and percentage of the maximum item score attained at posttest. The mean score increase was calculated as follows:

$$\Delta M = M_{post} - M_{pre}$$

where M_{post} and M_{pre} represent the mean posttest and pretest scores, respectively. The percentage of the maximum item score (P) attained at posttest was calculated as follows:

$$P = \frac{M_{post}}{16} \times 100\%$$

Individual improvement was measured using normalized gain (N-gain):

$$g_{N,i} = \frac{X_{post,i} - X_{pre,i}}{64 - X_{pre,i}}$$

where $X_{post,i}$ and $X_{pre,i}$ denote the posttest and pretest scores of student i, respectively, and 64 is the maximum possible test score. Following Hake (1998), N-gain values were classified as high ($g_N \geq 0.70$), moderate ($0.30 \leq g_N < 0.70$), or low ($g_N < 0.30$). These categories were used only for descriptive interpretation.

Between-group differences in improvement were examined using individual N-gain scores. Normality was assessed using the Lilliefors test and homogeneity of variance using an F test. Potential outliers were inspected using boxplots. After the normality and homogeneity assumptions were satisfied, mean N-gain scores were compared using a two-tailed pooled-variance independent-samples t test.

Hedges' g, a small-sample bias-corrected standardized mean difference, was also calculated to quantify the magnitude of the between-group difference (Lakens, 2013):

$$g = J\left(\frac{M_E - M_C}{S_p}\right)$$

where M_E and M_C represent the mean N-gain scores of the experimental and control groups, S_p is the pooled standard deviation, and J is the small-sample correction factor.

Because the groups showed different pretest scores, ANCOVA was additionally conducted as a sensitivity analysis to examine whether the between-group difference remained after accounting for initial performance. Posttest score was entered as the dependent variable, instructional group as the fixed factor, and pretest score as the covariate. Homogeneity of regression slopes was examined through the group-by-pretest interaction. Residual normality and homoscedasticity were examined using residual diagnostics. Standardized residuals and Cook's distance were inspected to identify potentially outlying or influential observations. Statistical significance was evaluated at $\alpha = .05$ for all inferential analyses.

3. RESULTS AND DISCUSSION

3.1. Results

The analysis focused on the pretest, posttest, and N-gain scores for mathematical critical thinking in the experimental and control groups. Table 2 presents the descriptive statistics for these measures.

Table 2. Descriptive Statistics for Mathematical Critical Thinking

Group	Measure	Mean	*SD*	Minimum	Maximum
Experimental	Pretest	12.25	7.59	4	30
(n = 16)	Posttest	39.88	11.28	24	58
	N-gain	0.548	0.175	0.333	0.887
Control	Pretest	8.17	3.52	1	14
(n = 24)	Posttest	31.25	7.84	22	52
	N-gain	0.416	0.130	0.236	0.765

Note. The maximum possible pretest and posttest score was 64.

As shown in Table 2, the mean posttest score was higher than the mean pretest score in both groups. In the experimental group, the mean score increased from 12.25 to 39.88, while in the control group, it increased from 8.17 to 31.25. The experimental group had a higher mean pretest score than the control group. Figure 2 displays the distributions of individual pretest and posttest mathematical critical thinking scores in both groups. Consistent with Table 2, the posttest score distributions were higher than the pretest distributions in both groups, while the experimental group also showed higher pretest scores than the control group.

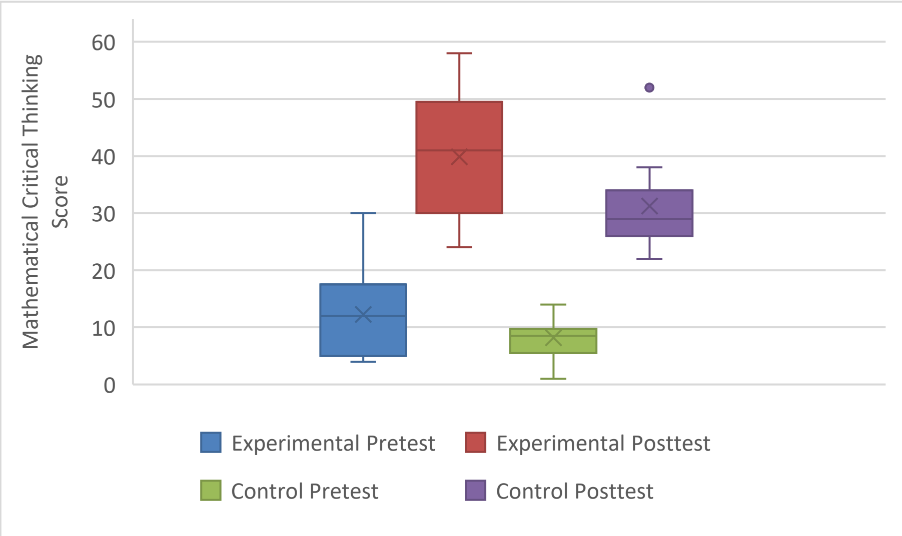


Figure 2. Pretest–Posttest Mathematical Critical Thinking Scores by Group

The experimental group obtained a mean N-gain of 0.548, compared with 0.416 in the control group. Based on the N-gain criteria, improvement in both groups was classified as moderate. Although both means fell within the same category, the experimental group had a higher mean N-gain by 0.132. An inferential test was subsequently conducted to examine the difference in mean N-gain between the two groups.

Item-level descriptive analyses were conducted to characterize the performance patterns of the two groups across interpretation, analysis, evaluation, and inference. The results are presented in Table 3.

Table 3. Posttest Percentage and Mean Score Increase by Indicator

Indicator	Group	M_{pre}	M_{post}	ΔM	P (%)
Interpretation	Experimental	5.375	9.125	3.750	57.03
	Control	4.333	10.750	6.417	67.19
Analysis	Experimental	4.313	11.375	7.063	71.09
	Control	0.917	11.250	10.333	70.31
Evaluation	Experimental	1.125	7.125	6.000	44.53
	Control	1.042	4.833	3.792	30.21
Inference	Experimental	1.438	12.250	10.813	76.56
	Control	1.875	4.417	2.542	27.60

Table 3 shows that the posttest score percentages (P) and mean score increases (ΔM) varied across the four indicators. In the experimental group, inference had the highest posttest percentage (76.56%), followed by analysis (71.09%), interpretation (57.03%), and evaluation (44.53%). Inference also showed the largest mean score increase, at 10.813 points.

In the control group, analysis had the highest posttest percentage (70.31%), followed by interpretation (67.19%), evaluation (30.21%), and inference (27.60%). Analysis also showed the largest mean score increase in this group, at 10.333 points.

The experimental group obtained higher posttest percentages and mean score increases than the control group for evaluation and inference. For analysis, the posttest percentages of the two groups were relatively similar, at 71.09% for the experimental group and 70.31% for the control group. However, the control group showed a larger mean score increase for this indicator. The control group had a mean pretest score of 0.917, compared with 4.313 in the experimental group. For interpretation, the control group obtained both a higher posttest percentage and a larger mean score increase than the experimental group.

Before comparing the mean N-gain scores of the two groups, the assumptions of normality and homogeneity of variance were examined. The results of the assumption tests are presented in Table 4.

Table 4. Results of Assumption Tests for N-gain Scores

Test	Group	Test Statistic	Critical Value	Conclusion
Lilliefors normality test	Experimental	$L = 0.170$	$L_{\text{critical}} = 0.213$	Normally distributed
Lilliefors normality test	Control	$L = 0.160$	$L_{\text{critical}} = 0.176$	Normally distributed
Homogeneity of variance	Both groups	$F = 1.813$	$F_{\text{critical}} = 2.130$	Homogeneous variances

For both groups, the calculated L values were smaller than their respective critical values, indicating that the N-gain scores were normally distributed. The homogeneity test yielded $F = 1.813$, which was smaller than $F_{\text{critical}} = 2.130$. Thus, the N-gain variances of the two groups were homogeneous.

Boxplot inspection identified two relatively high N-gain scores in the control group. Both values were retained because they represented valid observations within the possible N-gain range and were not attributable to data-entry errors.

After the normality and homogeneity assumptions were satisfied, the mean N-gain scores were compared using a pooled-variance independent-samples t test. The experimental group obtained a higher mean N-gain (0.548) than the control group (0.416). The difference was statistically significant, $t(38) = 2.731$, $p = .010$. Hedges' $g = 0.864$, corresponding to a standardized mean difference of approximately 0.86 pooled standard deviations. Thus, the N-gain analysis indicated greater mean improvement in the experimental group.

Because the experimental group had a higher mean pretest score, ANCOVA was conducted as a supplementary sensitivity analysis with posttest score as the dependent variable and pretest score as the covariate. The group-by-pretest interaction was not significant ($p = .749$), providing no evidence of unequal regression slopes across the two groups. The adjusted posttest means were 37.04 for the experimental group and 33.14 for the control group. After controlling for pretest performance, the between-group difference was not statistically significant, $F(1, 37) = 2.664$, $p = .111$. Standardized residuals did not exceed ±3, although four observations exceeded the 4/N screening value for Cook's distance; the largest Cook's distance was 0.193. These values were treated as screening flags rather than grounds for case deletion, and all observations were retained. Residual diagnostics indicated some departure from normality, while no marked heteroscedasticity was observed. The ANCOVA result was therefore interpreted as a supplementary sensitivity analysis rather than as the sole basis for inference.

3.2. Discussion

The findings provide mixed evidence regarding the between-group difference in mathematical critical thinking. The experimental group showed a significantly higher mean N-gain than the control group, with Hedges' $g = 0.864$, corresponding to approximately 0.86 pooled standard deviations. However, Hedges' g describes the magnitude of the same N-gain contrast; it does not provide the baseline-adjusted comparison obtained through ANCOVA. When pretest performance was taken into account, the adjusted posttest difference was not statistically significant. Taken together, the results indicate a possible advantage in observed improvement for the experimental group, but they do not provide conclusive evidence that e-worksheet-assisted PBL was superior to conventional expository instruction.

The experimental group's higher N-gain may be interpreted partly in terms of the structuring role of the e-worksheet within the PBL process. Hmelo-Silver (2004) describes structured supports in PBL as helping learners externalize problem solving, monitor their progress, and articulate and justify their thinking. Technology can also distribute some of these facilitation functions when direct facilitator support is limited. In the present study, guiding questions, spaces for developing solutions, and conclusion prompts may have played a comparable role by directing students' attention to successive stages of problem solving and prompting them to make parts of their thinking explicit. Broader evidence from inquiry-oriented instruction likewise indicates that student investigation is more productive when accompanied by appropriate guidance and that digital systems can carry part of this support (de Jong et al., 2023). Structured guidance is therefore one plausible interpretation of the observed N-gain pattern. However, the present outcome data do not show whether these particular worksheet features accounted for the between-group difference.

Although the experimental group showed higher descriptive performance on both evaluation and inference, the most pronounced separation occurred on the inference item, with posttest performance of 76.56% in the experimental group and 27.60% in the control group. In this study, the inference item required students to draw a conclusion supported by their analysis of the data. The e-worksheet also included an explicit section for drawing conclusions after students worked through the problem. This alignment is conceptually relevant because inference involves drawing reasonable conclusions from available evidence (Facione, 2015), and in statistical contexts it includes generating conclusions from an examination of the situation and the data (Koga, 2022). This task alignment offers one plausible interpretation of why the descriptive separation was especially marked on the inference item. However, because each indicator was represented by a single item and the indicator-level analyses were descriptive, this result should be read as a task-level descriptive pattern rather than as evidence of an indicator-specific treatment effect.

Performance on the evaluation item remained below half of the maximum score in both groups, at 44.53% in the experimental group and 30.21% in the control group. This pattern is consistent with evidence that procedural knowledge of mean, median, and mode does not necessarily extend to judging which measure is appropriate for a particular data set. Karaca and Ay (2025) found that eighth-grade students were more successful on operational tasks than on tasks requiring interpretation and selection of an appropriate measure of central tendency. Rosidah and Ikram (2021) similarly reported that university students often provided inadequate justifications and failed to consider relevant features such as outliers and data distribution. In the present study, the evaluation item likewise required students to assess the appropriateness of a choice, statement, or justification based on the data rather than merely perform a calculation. The experimental group's higher descriptive score therefore does not indicate that this broader difficulty was overcome. Although the e-worksheet provided opportunities to consider and justify solutions, performance on the evaluation item remained below half of the maximum score.

A different pattern emerged on the interpretation and analysis items. On interpretation, the control group reached a higher posttest percentage and showed a larger mean score increase than the experimental group. On analysis, posttest performance was nearly identical, at 71.09% in the experimental group and 70.31% in the control group. Although the control group showed a larger mean increase on this item, it also began from a lower pretest score. The raw increase should therefore be interpreted cautiously. These item-level patterns indicate that the experimental group's higher overall N-gain was not reflected uniformly across the assessed tasks. At a broader instructional level, de Jong et al. (2023) argue that instructional approaches should be considered in relation to the learning outcomes they are intended to support rather than treated as a simple hierarchy between inquiry-oriented and direct instruction. In the present study, e-worksheet-assisted PBL did not yield higher descriptive posttest performance on every assessed item.

The N-gain result is consistent in direction with closely related evidence from mathematics education, while broader PBL research provides a more qualified context for interpreting that pattern. Safaatullah et al. (2025) reported higher posttest mathematical critical thinking performance among eighth-grade students receiving Google Sites-assisted PBL than among those receiving conventional instruction. Their study, however, focused on the Pythagorean theorem and used a posttest-only design. In contrast, the present study examined measures of central tendency and included pretest measurement, allowing a supplementary baseline-adjusted comparison through ANCOVA. The forms of digital support also differed. Google Sites was used to provide learning materials, videos, problem illustrations, and practice activities, whereas the present e-worksheet was designed to structure successive stages of problem solving through guiding questions, solution spaces, and conclusion prompts. At a broader level, Lu et al. (2025) found an overall advantage of PBL for critical thinking, but also substantial heterogeneity across the included studies, with significant variation associated with teaching method and subject. The present findings therefore align with the generally positive direction of this literature while also showing that the strength of the between-group evidence can depend on how baseline performance is taken into account.

The standardized N-gain difference should be interpreted as a measure of magnitude rather than as independent confirmation of an instructional advantage. Hedges' $g = 0.864$ indicates that the group means on N-gain differed by approximately 0.86 pooled standard deviations. Because this effect size was calculated from the same N-gain contrast examined by the t test, it does not provide the baseline-adjusted posttest comparison obtained from ANCOVA. Under conventional benchmarks, a standardized difference of 0.864 would be described as large. However, such category thresholds are arbitrary and should not be treated as direct indicators of practical importance (Lakens, 2013). The supplementary ANCOVA addressed a different comparison by estimating the posttest group difference after accounting for pretest scores, and this adjusted difference was not statistically significant. Accordingly, the significant N-gain difference describes the observed pattern of improvement, but it does not support an unqualified claim that e-worksheet-assisted PBL was superior after baseline performance was taken into account.

Beyond the overall between-group comparison, the findings refine how digitally supported PBL may be understood in statistics instruction. Digital-based approaches are common among integrated forms of PBL in K–12 mathematics, whereas data, probability, and statistics remain comparatively underrepresented (Ayari et al., 2025). The present results add a qualification to this literature. Higher descriptive performance under e-worksheet-assisted PBL did not occur on every assessed item, suggesting that the presence of digital support alone is not sufficient for interpreting its instructional contribution. In this study, the e-worksheet was designed to guide students through successive stages of problem solving, yet the item-level patterns varied across the assessed task demands. This variation raises the possibility that the value of structured guidance depends partly on how its prompts and activities align with the statistical judgments required by the target task.

This interpretation remains provisional because the study did not include process data showing how students actually used the guidance.

Several limitations constrain the interpretation and generalizability of these findings. The study involved only two intact classes from a single school, with one class representing each instructional condition. Class-level characteristics therefore could not be separated fully from the instructional condition. The small sample, comprising 16 students in the experimental group and 24 in the control group, limits the precision of the between-group estimates and the extent to which the findings can be generalized beyond this setting. Although the classes had similar prior mathematics report-card averages, the experimental group began with a higher mean pretest score, indicating residual baseline imbalance. Unmeasured student-level factors, such as learning motivation and familiarity with digital learning tools, may also have contributed to the observed performance patterns. Both groups were taught by the same researcher, which reduced between-teacher variation. However, treatment fidelity was not independently observed or quantified, so unplanned differences in implementation cannot be ruled out. The intervention was limited to three instructional sessions, restricting conclusions to relatively short-term outcomes. In addition, PBL and the e-worksheet were implemented as a single instructional configuration, so their separate contributions cannot be determined from the present data. Finally, the ANCOVA showed some departure from residual normality, which adds further caution to the baseline-adjusted comparison.

The indicator-level findings have additional measurement limitations. Each indicator was represented by a single test item, so the observed patterns are tied to the particular tasks used. An empirical reliability estimate was not available for the four-item test, limiting the extent to which the consistency of the resulting scores can be evaluated. Using the same instrument at pretest and posttest may also have introduced some familiarity with the items.

4. CONCLUSION

This study provides mixed evidence regarding the between-group difference in mathematical critical thinking on measures of central tendency. The experimental group showed greater observed improvement based on N-gain, whereas the baseline-adjusted posttest difference examined through ANCOVA was not statistically significant. The findings therefore suggest a possible advantage for e-worksheet-assisted PBL in observed improvement, but they do not provide conclusive evidence of superiority over conventional expository instruction.

Theoretically, the study contributes a more qualified view of digitally supported PBL in school statistics. The findings suggest that the contribution of structured digital guidance may depend partly on how its prompts and activities align with the thinking demands of particular statistical tasks. This task-alignment perspective shifts attention from the presence of digital support itself to the relationship among e-worksheet guidance, PBL activities, and the statistical judgments students are expected to make.

Practically, e-worksheets used within PBL can be designed as structured supports rather than merely as digital repositories of learning materials. Guiding questions, spaces for developing solutions, and conclusion prompts can be incorporated to structure opportunities for students to interpret information, evaluate statistical choices, and draw conclusions from data. Particular attention may be warranted for tasks requiring evaluation, given the relatively low performance on the evaluation item in both groups.

These conclusions are limited by the use of two intact classes from a single school, a small sample, baseline imbalance between the groups, and an intervention lasting only three instructional sessions. Treatment fidelity was not independently quantified, and no empirical reliability estimate was available for the four-item test. In addition, PBL and the e-worksheet were implemented as a single instructional configuration, so their separate contributions cannot be determined. Future studies should involve multiple schools and classes, longer interventions, and multiple items for each critical thinking indicator. Process-oriented and mixed-methods research could examine how students actually use digital guidance while working with statistical tasks. Studies across other mathematical topics and educational levels could examine whether the present patterns extend beyond measures of central tendency and the educational setting investigated here.

REFERENCES

Arends, R. I. (2012). *Learning to teach* (9th ed.). McGraw-Hill.

Ayari, M. A., Sellami, A., Santhosh, M. E., Naji, K. K., Al-Ali, A., & Al-Hazbi, S. M. A. (2025). From problems to performance: A systematic review of problem-based learning in K-12 mathematics. *Frontiers in Education,* 10, 1731307. https://doi.org/10.3389/feduc.2025.1731307

de Jong, T., Lazonder, A. W., Chinn, C. A., Fischer, F., Gobert, J., Hmelo-Silver, C. E., Koedinger, K. R., Krajcik, J. S., Kyza, E. A., Linn, M. C., Pedaste, M., Scheiter, K., & Zacharia, Z. C. (2023). Let's talk evidence–The case for combining inquiry-based and direct instruction. *Educational Research Review*, *39*, 100536. https://doi.org/10.1016/j.edurev.2023.100536

Facione, P. A. (2015). *Critical thinking: What it is and why it counts* (2015 update). Measured Reasons LLC.

Febriani, I. R. F., Rahaju, E. B., Ekawati, R., & Shodikin, A. (2024). Statistics flip-worksheet: The key to improve students' critical thinking skill. *Journal of Mathematical Pedagogy*, *5*(2), 76–90. https://doi.org/10.26740/jomp.v5n2.p76-90

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, *66*(1), 64–74. https://doi.org/10.1119/1.18809

Hmelo-Silver, C. E. (2004). Problem-based learning: What and how do students learn? *Educational Psychology Review*, *16*(3), 235–266. https://doi.org/10.1023/B:EDPR.0000034022.16470.f3

Karaca, Ş. A., & Ay, Z. S. (2025). Investigation of eighth grade students’ performance on tasks involving statistical thinking about measures of central tendency. *Participatory Educational Research*, *12*(1), 18–42. https://doi.org/10.17275/per.25.2.12.1

Koga, S. (2022). Characteristics of statistical literacy skills from the perspective of critical thinking. *Teaching Statistics*, *44*(2), 59–67. https://doi.org/10.1111/test.12302

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863. https://doi.org/10.3389/fpsyg.2013.00863

Lu, L., Mustakim, S. S., & Muhamad, M. M. (2025). A meta-analysis of the effectiveness of problem-based learning on critical thinking. *European Journal of Educational Research*, *14*(3), 789–804. https://doi.org/10.12973/eu-jer.14.3.789

Monteleone, C., Miller, J., & Warren, E. (2023). Conceptualising critical mathematical thinking in young students. *Mathematics Education Research Journal*, *35*(2), 339–359. https://doi.org/10.1007/s13394-023-00445-1

Muerza, V., Gargallo, P., Salvador, M., & Turón, A. (2024). Impact of problem-based learning on the perception, understanding, and application of statistical concepts in business administration and management students. *Sustainability, 16*, 1591. https://doi.org/10.3390/su16041591

Muslimin, M., Handayanto, S. K., & Sari, I. N. (2024). Influence of e-scaffolding in problem-based learning on students' HOTS in standing wave. *International Journal of Interactive Mobile Technologies*, *18*(9), 100–116. https://doi.org/10.3991/ijim.v18i09.47957

Rosidah, R., & Ikram, F. Z. (2021). Measure of central tendency: Undergraduate students' error in decision-making perspective. *International Journal of Education*, *14*(1), 39–47. https://doi.org/10.17509/ije.v14i1.29408

Safaatullah, M. F., Pangestu, A. M., & Amidi, A. (2025). Students' mathematical critical thinking ability reviewed from learning motivation in problem-based learning model assisted by Google Sites. *Unnes Journal of Mathematics Education*, *14*(1), 22–35. https://doi.org/10.15294/ujme.v14i1.22151

Tempelaar, D., Bátori, A., & Giesbers, B. (2024). Understanding self-regulation strategies in problem-based learning through dispositional learning analytics. *Frontiers in Education, 9*, 1382771. https://doi.org/10.3389/feduc.2024.1382771