*Edumatica: Jurnal Pendidikan Matematika*
*Volume 16 Number 1 April 2026*

Submited : 27 June 2025
Revised : 15 August 2025
Accepted : 28 December 2025

DOI: https://doi.org/10.22437/edumatica.vl6i1.46293

# The Effect of Worked-Examples on Students' Learning Outcomes in Complex Mathematics Materials

**Rodiyatul Hafidhoh[1], Cecep Anwar Hadi Firdos Santosa[2*], Anwar Mutaqin[3]**
[1,2,3]Universitas Sultan Ageng Tirtayasa, Indonesia
E-mail*: cecepanwar@untirta.ac.id

## Abstract

This study aimed to examine the effect of the worked-example strategy on students' learning outcomes in complex mathematics across three time points (pre-test, post-test, delayed-test), and to analyze whether this effect differed based on students' gender and initial ability levels. A quantitative one-group within-subjects quasi-experimental design was used with 31 eighth-grade students from SMPN 1 Pulosari. Data were collected through validated and reliable test instruments and analyzed using Rasch modeling and inferential statistics. Results showed that the strategy significantly improved learning outcomes ($p<0.05$), particularly from pre-test to post-test. Although delayed-test scores slightly declined, performance remained high. Moreover, initial differences based on ability levels were observed in the pre-test but they were no longer significant ($p>0.05$) in the post-test and delayed-test results, suggesting that the strategy helped lower-ability students catch up, reducing learning gaps. Additionally, the study found no significant differences in learning outcomes based on gender ($p > 0.05$), nor any interaction between gender or ability level and the effectiveness of the worked-example strategy ($p>0.05$). This study demonstrates the worked example strategy's effectiveness across time, abilities, and genders in reducing cognitive load and strengthening students' conceptual understanding. Future studies may compare types, involve varied levels, and use more rigorous designs.

**Keywords**: cognitive load; complex mathematics; learning outcomes; worked-example

## INTRODUCTION

Mathematics is one of the lessons that is often considered difficult by many students, resulting in unsatisfactory learning outcomes (Ayu et al., 2024; Setiawan et al., 2016). In addition to students' mindset towards mathematics, complex mathematics material is also sometimes a trigger for low student learning outcomes. Learning outcomes are an important indicator in evaluating the success of learning. Learning outcomes reflect the extent to which an individual has mastered the material through the results of assessments, serving as an indicator of learning outcomes (Nainggolan & Pasaribu, 2021). In this study, learning outcomes include two main types of knowledge (procedural knowledge and conceptual knowledge**)**. One of the factors influencing students' learning outcomes is the level of complexity of the material being studied, as complex material can increase the difficulty of understanding concepts and solving mathematical problems.

Complex mathematics material refers to concepts and topics in mathematics that have a high level of difficulty or complexity and involve many interactions between various elements or components. Element interactivity refers to how various learning components, such as numbers, symbols, and procedures are interconnected and need to be processed at the same time within working memory (Adeniji & Baker, 2023; Chen, 2015; Sweller, 2011). The level of element interactivity can be identified by estimating the number of interacting elements in a learning material. The more elements that must be processed simultaneously, the higher the level of interactivity and cognitive load that students may experience, especially intrinsic cognitive load (Adeniji & Baker, 2023; Chen, 2015). This can affect students' ability to process information effectively.

In addition to element interactivity, complex mathematics material is typically characterized by the need for deep procedural and conceptual understanding. Procedural understanding involves the ability to follow steps to solve problems, while conceptual understanding relates to principles and the relationships between mathematical concepts. In learning complex mathematics, these characteristics high element interactivity and the need for deep procedural and conceptual understanding can increase the difficulty for students, especially those with low ability. Complex material with various abstract elements is often a challenge for students, resulting in high cognitive load, particularly extrinsic cognitive load (Richardo & Cahdriyana, 2021; Santosa & Filiz, 2025).

According to Santosa et al (2018) Increased cognitive load occurs due to the limitations of working memory in handling information, making it vulnerable to overload, which ultimately inhibits learning. In this case, when the amount of information that must be processed by students exceeds their working memory capacity, students will have difficulty processing and remembering information (Sholihah, 2022). Thus, success in learning is highly dependent on the ability of one's memory to process and store the information received.

Through learning, individuals not only accumulate new knowledge but also utilize memory skills to store, recall, and apply the information, so that it can be implemented in the future. According to Adeniji et al (2018) The learning process does not stop at mastering information at that time, but also the ability of students to remember and apply that knowledge in different contexts in the future. Therefore, it is important for teachers to design various efforts to develop good mathematics learning methods in order to support student understanding during the learning process, to support long-term knowledge retention, especially for complex material.

One of the effective learning strategies to achieve these goals can refer to a cognitive load theory found by Sweller, namely Cognitive Load Theory (CLT), which is a theory that deals with how to design instructions and deliver information efficiently, taking into account the limited ability of working memory. According to Sweller (2011) CLT is a learning theory based on knowledge of human cognition. CLT is able to optimize instructional design to manage working memory limitations and utilize long-term memory, thus improving learning and performance on complex tasks (Sholihah, 2022; Sweller, 1988). By optimizing instructional design and reducing unnecessary cognitive load, CLT can directly support the enhancement of students' learning outcomes. Specifically, it helps students master complex mathematics concepts, follow procedural steps accurately, and develop conceptual understanding, which are the main indicators of learning outcomes in this study

Three sources of cognitive load can be identified according to Sweller (2011), namely intrinsic cognitive load, extrinsic, and constructive cognitive load. Intrinsic cognitive load is directly related to the ability of students to understand the information received (Ratnasari, 2023), this cognitive load cannot be changed because it depends on the complexity of the material and the material taught, and its elements (Irwansyah & Retnowati, 2019; Richardo & Cahdriyana, 2021; Santosa et al., 2018; Sweller, 2011). While extrinsic cognitive load is related to the way the presentation material, such as methods and strategies applied by teachers in the classroom and can still be controlled (Afidah, 2020; Ratnasari, 2023; Richardo & Cahdriyana, 2021; Santosa et al., 2019; Sweller, 2011). Furthermore, the constructive cognitive load is where this load is related to the learning process, especially in fostering knowledge sche ma construction and successful problem solving (Irwansyah & Retnowati, 2019; Santosa et al., 2019).

Of the three, cognitive load that can inhibit the learning process is extrinsic cognitive load (Muryanto, 2020), this can be manipulated by the preparation of methods, strategies, or learning designs to increase the effectiveness of learning (Anisa & Endah Retnowati, 2024). One of the strategies that can be used to design learning is the worked-example. This strategy has been proven to encourage knowledge transfer in mathematics learning, both individually and collaboratively (Barbieri et al., 2023; Retnowati et al., 2010; Sholihah, 2022), in addition to supporting the limitations of students' initial abilities and reducing external cognitive load, this worked-example strategy is very effective in learning (Sweller et al :2011; Irwansyah & Retnowati, 2019).

A worked-example is a strategy that presents students with fully solved examples, showing each step of the solution in detail. A worked-example demonstrates problem solving by presenting the solution step by step, starting from problem formulation, continuing with the solving process, and ending with the final answer (Hoogerheide, 2014; Santosa et al., 2022). Showing these steps is intended to guide students in learning how to reach solutions more effectively and with greater understanding (Atkinson et al., 2000; Intan & Rosyid, 2020)

The worked-example strategy can effectively reduce cognitive load, which makes each learning step less challenging, thus improving learning outcomes, especially in problem solving (Chen et al., 2023; Hoogerheide, 2014; Irwansyah & Retnowati, 2019; Santosa et al., 2018). This supports the application of worked-example in learning that can affect learning outcomes. However, further research is needed on how the worked example strategy can influence learning outcomes in complex mathematics contexts, especially in terms of its short-term and long-term effects.

Beyond cognitive factors, learning outcomes can also be influenced by gender and students' ability levels. Several studies indicate that gender stereotypes still play a role in mathematics learning outcomes (Ayebale et al., 2020; Yulianto et al., 2025), while others report that male and female students now perform at relatively similar levels (Adeniji & Baker, 2023; Hyde, 2014). So the influence of gender is still inconclusive. On the other hand, students' ability level more consistently affects learning outcomes. High-ability students tend to have stronger cognitive schemas (Adeniji & Baker, 2023; Kalyuga & Renkl, 2010; Sweller, 1988), whereas low-ability students require more instructional support (Alreshidi, 2021; Rohman & Retnowati, 2018). It is therefore essential to examine how these strategies can be adapted to suit students with varying ability levels.

Several studies have explored how effective worked-example strategies are in supporting learning. Adeniji & Baker (2023) found that worked-example significantly improved short-term learning outcomes (pre-test to post-test), though its effect on long-term outcomes (post-test to delay-test) was weaker. It was more beneficial for low-ability students, and no significant differences were found based on gender. However, this study was conducted during the COVID-19 pandemic, limiting direct interaction between researchers and students. The present study revisits the worked-example strategy under more stable post-pandemic learning conditions, allowing more intensive interaction and observation to generate more accurate data on its effects on complex mathematics learning outcomes.

Similarly, Alreshidi (2021) reported that example-problem pairs improved students' mathematics achievements, especially for students with sufficient prerequisite knowledge, but this study focused only on male students due to gender-segregated schooling in Saudi Arabia. The current study expands the context by including both genders and analyzing the effects of ability and gender on learning outcomes.

 

Chen et al. (2023) showed that worked-examples positively affected retention and knowledge transfer while reducing cognitive load, but retention was measured only one week after learning. In this study, a delay test conducted three weeks after the post-test allows for a more comprehensive assessment of long-term learning outcomes. Despite these findings, research examining how worked-example strategies interact with gender and student ability in affecting both short-term and long-term learning outcomes is still limited in Indonesia. This represents the research gap that this study aims to address.

Based on this background, this study was conducted to examine the effectiveness of the worked-example strategy in improving student learning outcomes in complex mathematics materials at SMPN 1 Pulosari. This research focused on several main questions: (1) Does the worked-example strategy have an effect on student learning outcomes at three different time points? (2) Are there differences in learning outcomes based on student ability levels when applying the worked-example strategy? (3) Are there differences in learning outcomes based on student gender when applying the worked-example strategy? (4) Does gender or student ability level interact with the effectiveness of worked-example?

## METHOD

This study used a quantitative approach with a one-group, within-subject, quasi-experimental design, involving one class of students tested at three different time points, namely the pre-test, post-test, and delayed-test. The one-group within-subjects quasi-experimental design was selected because it allows for measuring changes in the same group of participants over multiple time points (pre-test, post-test, and delayed-test). This approach reduces variability caused by individual differences, as each student serves as their own control. It was also the most feasible design given the limited number of classes available for research in the school, as the school only permitted the use of one intact class. Despite the absence of a control group, this design is appropriate for examining the effectiveness of the worked-example strategy over time and provides valuable preliminary evidence that can be further tested using more rigorous designs in future research. The research was conducted at SMP Negeri 1 Pulosari, Pandeglang, with a sample of 31 eighth-grade students (16 females and 15 males). This study aims to measure the effect of the worked example strategy on complex math learning outcomes at three different measurement times, namely before treatment (pre-test), after treatment (post-test), and three weeks after the post-test (delayed-test).

The mathematics material used in this study was the topic of functions for eighth-grade junior high school students, categorized as complex material. This selection was based on its high element interactivity, requiring students to understand relationships between variables, use various representations (tables, graphs, and equations), and engage in abstract and symbolic mathematical reasoning. The complexity of this topic was further supported by the subject teacher's observation that functions often present challenges, particularly for students with weak foundational mathematical skills. Therefore, functions were considered likely to impose a high cognitive load, requiring an appropriate instructional strategy, such as the worked-example approach, to help students process information effectively.

The instrument used in this study was a test designed to measure the effect of the worked-example strategy on students' learning outcomes. This test instrument was designed to be used in the pre-test, post-test, and delayed-test. It consisted of five essay-type questions: two items assessed conceptual knowledge (e.g., identifying whether a relation is a function and explaining examples of functions in daily life), and three items assessed procedural knowledge (e.g., representing functions in various forms, determining the values and graphs of functions, and constructing linear function formulas from given information). All items were designed to measure complex mathematics learning outcomes, particularly problem-solving in function-related problems with high element interactivity, which require the integration of multiple concepts and solution steps.

Experts validated the test instrument to confirm its accuracy, clarity, and consistency with the targeted learning objectives. The validators included the first and second academic supervisors, who are experts in learning and instructional design, as well as an experienced mathematics teacher. They reviewed the instrument using a validation sheet covering aspects such as content relevance, clarity of instructions, appropriateness of difficulty level, and the balance between conceptual and procedural

knowledge. Feedback from the validators was used to make revisions, ensuring that the instrument was both valid and reliable before being implemented in the study.

With a validated and reliable instrument in place, the study proceeded to the learning implementation stage, which lasted for three meetings. During this phase, all students in the class received the same treatment, namely participating in the learning process using the worked-example strategy. This strategy was applied consistently in each meeting to help students understand complex mathematics material through examples accompanied by step-by-step solution tracing.

Each meeting began with an initial learning phase, including an aperception and elicitation of students' prior knowledge through brief question-and-answer sessions. The teacher then provided a concise explanation of the basic concepts of the topic to build initial conceptual understanding and ensure that students had a sufficient foundation before progressing to more complex material.

In the main learning phase, the teacher presented the core material along with fully solved example problems (worked examples) on the board. These examples included sequential solution steps with accompanying verbal explanations. Students were instructed to carefully observe the examples while taking notes and highlighting the key steps.

Next, students were guided to review the material independently and were given opportunities to ask the teacher about parts they did not yet understand. This activity aimed to strengthen both students' conceptual and procedural understanding through tracing and repetition. Following this, students were provided with acquisition sheets in an example–problem pairs format, containing several solved example problems each followed by similar practice problems. These sheets were intended to help students develop conceptual schemas and apply their understanding, preparing them for the post-test and delayed-test.

After the three meetings, a post-test was conducted to measure students' learning outcomes following the intervention. Three weeks later, a delayed-test was administered to assess students' long-term retention of the material. The post-test and delayed-test items were designed to be similar to those on the pre-test.

The collected data were then analyzed using the SOLO Taxonomy to evaluate students' conceptual and procedural understanding. Subsequently, Rasch analysis was performed with the assistance of Winsteps software to estimate students' abilities, determine item difficulty, and assess data-model fit, providing insight into the interaction between item characteristics and student performance.

Following this, inferential statistical analyses were conducted. Preliminary assumption tests, including normality (Shapiro-Wilk) and homogeneity (Levene), were carried out as prerequisites. Paired sample t-tests were used to compare mean scores between pre-test and post-test, as well as between post-test and delayed-test. Independent sample t-tests were applied to examine differences in learning outcomes based on students' ability levels and gender. Repeated measures ANOVA was employed to analyze differences across the three measurement points, and mixed-design ANOVA was used to test potential interactions between within-subject factors (time points) and between-subject factors (ability or gender). This analytical procedure allowed for a comprehensive evaluation of the effectiveness of the worked-example strategy over time and across different student groups.

## RESULTS

The research results consisted of pre-test, post-test, and delayed-test data to measure students' learning outcomes on complex math materials with the Worked-example strategy at three different time points. Student responses at the three time points were assessed with reference to the SOLO Taxonomy and continued with Rasch analysis as well as inferential. Unlike classical statistical approaches, the Rasch model has its own advantages (Rahim & Haryanto, 2021). One of them is that a student's correct answer is not only determined by his or her ability, but also by the nature and difficulty of the question (Adi et al., 2022; Muhtarom, 2024). In other words, the correct answer of a student is not only determined by his/her ability, but also by the way the question is designed and the difficulty level. The results of the Rasch analysis are presented in Table 1:

 

Table 1 Summary Statistics of Rasch for Item and Response Estimation

| Test | Rasch separation indeks (I) | Rasch separation indeks (S) | Infit (I) | Infit (S) | Out fit (I) | Out fit (S) | Reliability (I) | Reliability (S) |
|---|---|---|---|---|---|---|---|---|
| Pre-test | 2.59 | 1.09 | 0.99 | 0.89 | 0.70 | 0.70 | 0.87 | 0.54 |
| Post-test | 5.09 | 3.35 | 1.02 | 0.99 | 0.98 | 0.98 | 0.96 | 0.92 |
| Delayed -test | 4.72 | 3.43 | 0.98 | 1.00 | 0.95 | 0.95 | 0.96 | 0.92 |

Table 1 displays the summary results of the Rasch analysis for item and student response estimates across the pre-test, post-test, and delayed test. The table reports separation indices for items and students, infit and outfit statistics, and reliability values. The findings indicate that the item separation index improved from the pre-test to the later tests, suggesting that the instrument became more effective in distinguishing the difficulty levels of the questions. Similarly, the student separation index also increased, indicating the instrument's ability to differentiate student ability levels after treatment. The infit and outfit values are within the ideal range (0.5-1.5), indicating the consistency of student responses to the model. The reliability of the instrument was very high in the post-test and delayed-test (I = 0.96; S = 0.92), and remained acceptable in the pre-test (I = 0.87; S = 0.54), so the instrument was considered reliable for measuring student learning outcomes. Student reliability of more than 0.5 indicates that there is more than one level of student ability (Adeniji, 2023). Students were categorized into two levels of ability, low and high, based on their performance on the pre-test. This finding supports further analysis of the effect of the worked-example strategy on learning outcomes based on students' skill levels.

To see the difficulty level of the questions and the level of student ability, it can be seen from the output of the Wright Map Item & Person Rasch table.

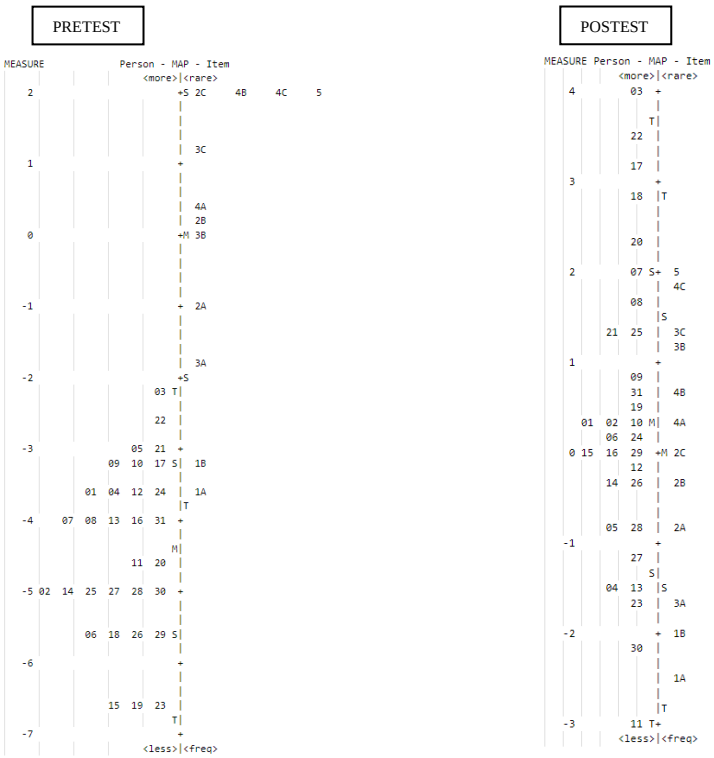


Figure 1 Output of the Wright Map Item & Person Rasch table

Figure 1 shows that in the pre-test, the majority of students were below 0 logits, while many items were above, indicating the questions were classified as difficult. After the treatment, in the post-test, the distribution of students shifted upwards, indicating an improvement in ability. The items were also evenly distributed along the logit scale, indicating a match between the difficulty level of the questions and the students' abilities. This suggests that the instrument effectively differentiates student ability and supports improved learning outcomes.

Next, further analysis was conducted to answer various research questions related to the effect of the worked-example strategy on student learning outcomes, as well as its relationship with gender and ability level. This analysis included inferential statistical testing, such as t-test, ANOVA, and Mixed Design ANOVA, to examine the significance of differences and interactions that occurred in various measurement conditions.

**Does the worked-example strategy affect student learning outcomes at three different time points?**

Before conducting inferential statistical tests, a descriptive analysis of student learning outcomes at three stages of measurement is presented. Table 2 presents descriptive statistics of students' scores at three measurement points: pre-test, post-test, and delayed-test. For each time point, the table displays the number of participants (N), minimum score (Min), maximum score (Max), mean score (Mean), and standard deviation (Std. Dev.). The mean score increased substantially from the pre-test (9.58) to the post-test (56.45), indicating a marked improvement after the intervention, and then slightly declined in the delayed-test (47.87), showing some decrease in retention over time. The standard deviation values suggest that score variability was low at the pre-test (5.46) but became larger in the post-test (21.01) and delayed-test (21.59), indicating greater differences in student performance after the treatment. To determine whether the difference is significant, inferential statistical tests need to be conducted. Before that, a prerequisite test was carried out in the form of a normality test to determine the appropriate type of test. The normality test using Shapiro-Wilk showed that the data at all three time points were normally distributed ($p > 0.05$), so it could be continued with the parametric test, namely the paired sample t-test.

Table 2 Descriptive Statistics of Students' Pre-test, Post-test, and Delayed-test

| Three points in time | N | Min | Max | Mean | Std. Dev |
|---|---|---|---|---|---|
| Pre-test | 31 | 2 | 23 | 9,58 | 5,46 |
| Post-test | 31 | 13 | 92 | 56,45 | 21,01 |
| Delayed-test | 31 | 6 | 90 | 47,87 | 21,59 |

Table 3 presents the results of the paired sample t-test comparing students' scores between the pre-test and post-test (Pair 1), and between the post-test and delayed-test (Pair 2). Based on the paired sample t-test results presented in table 3, there is a significant difference between the pre-test and post-test ($t(30) = -13.35$, $p = 0.000$), as well as between the post-test and delayed-test ($t(30) = 12.31$, $p = 0.000$). The p (Sig.) value <0.05 indicates that the changes between each time point were significant, indicating that the intervention implemented in this study had an effect on students' learning outcomes. Furthermore, to see the overall effect of the strategy and monitor the pattern of changes in learning outcomes, Repeated Measures ANOVA analysis was conducted.

Table 3 Results of Paired Sample T-Test

| | | t | df | Sig. (2-tailed) |
|---|---|---|---|---|
| Pair 1 | Pretest-Postest | -13,35 | 30 | 0,000 |
| Pair 2 | Postest-Delaytest | 12,31 | 30 | 0,000 |

Before conducting the Repeated Measures ANOVA analysis, a sphericity assumption test was conducted to ensure the similarity of variance differences between measurement times (pre-test, post-test, delayed-test). This test uses Mauchly's Test of Sphericity. If the significance value is <0.05, then

the assumption is not met, and the analysis continues with Greenhouse-Geisser correction. Conversely, if $p > 0.05$, the F value in the Sphericity Assumed row was used.

Based on the results of Mauchly's Test of Sphericity on Table 4, since the (Sig.) value of $0.00 < 0.05$, the assumption of sphericity is considered not met, so the ANOVA analysis must use the correction value (Greenhouse-Geisser) to correct the degrees of freedom.

Table 4 Test of Sphericity (Mauchly's Test Results)

| Within-Subjects Effect | Mauchly's W | Approx. Chi-Square | df | Sig. |
|---|---|---|---|---|
| Waktu | 0,57 | 16,13 | 2 | 0,000 |

Table 5 presents the results of the Repeated Measures ANOVA with the Greenhouse–Geisser correction, applied because Mauchly's Test indicated that the sphericity assumption was violated. The results (Sig. = $0.000 < 0.05$) show significant differences in students' mean learning outcomes across the pre-test, post-test, and delayed test. To identify where these differences occurred, a post-hoc Pairwise Comparison with Bonferroni correction was conducted.

Table 5 Repeated Measures ANOVA Results Using Greenhouse-Geisser Correction

| Tests of Within-Subjects Effects | | | | |
|---|---|---|---|---|
| | df | Mean Square | F | Sig. |
| Greenhouse-Geisser | 1,402 | 27543,651 | 130,273 | 0,000 |

Table 6 presents the results of the Post Hoc Pairwise Comparisons using the Bonferroni correction to compare the mean student learning outcomes scores across each pair of measurement times (pre-test, post-test, and delayed-test*)*. Based on table 6, there were significant differences between all pairs of measurement times (Sig. = 0.000). Learning outcomes increased sharply from pre-test to post-test (difference = 46.87), remained higher in the delayed-test than the pre-test (difference = 38.29), but decreased from post-test to delayed-test (difference = 8.58). This shows that despite the decrease after the time lag, students' learning outcomes remained better than before the treatment.

Table 6 Results of Post Hoc Pairwise Comparisons with Bonferroni Correction

| Pairwise Comparison Test | | | | | |
|---|---|---|---|---|---|
| Time (I) | Time (J) | Mean Difference (I-J) | Std. Error | Sig. (p) | 95% Confidence Interval |
| Pre-test | Post-test | -46.871 | 3.511 | 0.000 | (-55.775, -37.967) |
| Pre-test | Delayed-test | -38.290 | 3.609 | 0.000 | (-47.442, -29.139) |
| Post-test | Delayed-test | 8.581 | 1.824 | 0.000 | (3.954, 13.207) |

**Are there differences in learning outcomes based on students' ability levels when applying the worked-example strategy?**

Before inferential statistical tests were conducted, descriptive analysis was presented to provide an initial picture of student learning outcomes based on ability levels at three stages of measurement.

Table 7 presents descriptive statistics of students' learning outcomes based on their ability levels (high and low) at three measurement stages: pre-test, post-test, and delayed-test. Based on table 7, high-ability students showed higher average learning outcomes than low-ability students at all measurement times. This indicates that initial ability level can affect learning outcomes, both in the short and long term. To test the significance of these differences, an inferential test was conducted by first ensuring the assumptions of normality and homogeneity were met.

Table 7 Descriptive Data of Students Based on Ability Level

| Ability level | N | Min | Max | Mean | Std. Dev |
|---|---|---|---|---|---|
| High Pre-test | 16 | 10 | 23 | 13,92 | 3,75 |
| Low Pre-test | 15 | 2 | 8 | 4,93 | 1,98 |
| High Post-test | 16 | 29 | 92 | 62,68 | 19,83 |
| Low Post-test | 15 | 13 | 85 | 49,80 | 20,80 |
| High Delayed-test | 16 | 13 | 91 | 54,62 | 21,06 |
| Low Delayed-test | 15 | 6 | 79 | 40,66 | 20,39 |

Before conducting inferential tests to determine the statistical significance of these differences, the assumptions of normality and homogeneity were checked. Normality test results using Shapiro-Wilk showed that all data in the pre-test, post-test, and delayed-test were normally distributed ($p > 0.05$). The homogeneity test using Levene also showed that the variance between groups was homogeneous ($p > 0.05$). Thus, the data met the requirements for the next parametric test, namely the Independent Sample t-test.

Table 8 reports the Independent Sample t-test results comparing the learning outcomes of students with high and low ability across three assessments (pre-test, post-test, and delayed test). The pre-test results show a significant difference ($p = 0.000$) with a mean gap of 9.00. In contrast, the post-test ($p = 0.088$) and delayed test ($p = 0.071$) did not yield statistically significant differences, despite mean gaps of 12.89 and 13.96, respectively. These findings suggest that after the intervention, the performance gap between high- and low-ability students became more balanced.

Table 8 Independent Sample T-Test Results Based on Students' Ability Level

| Three points in time | t | df | Sig. (2-tailed) | Mean Difference |
|---|---|---|---|---|
| Pre-test | 8,27 | 29 | 0,000 | 9,00 |
| Post-test | 1,77 | 29 | 0,088 | 12,89 |
| Delayed-test | 1,87 | 29 | 0,071 | 13,96 |

Table 8 reports the Independent Sample t-test results comparing the learning outcomes of students with high and low ability across three assessments (pre-test, post-test, and delayed test). The pre-test results show a significant difference ($p = 0.000$) with a mean gap of 9.00. In contrast, the post-test ($p = 0.088$) and delayed test ($p = 0.071$) did not yield statistically significant differences, despite mean gaps of 12.89 and 13.96, respectively. These findings suggest that after the intervention, the performance gap between high- and low-ability students became more balanced.

**Is there a difference in learning outcomes based on student gender when applying the Worked-example strategy?**

Before inferential statistical tests are conducted, descriptive analysis is presented to provide an initial picture of student learning achievements based on ability levels at three stages of measurement. Table 9 shows the difference in average learning outcomes between male and female students at all three measurement times. Female students excel in the pre-test, while male students show higher averages in the post-test and delayed-test. To determine whether this difference is statistically significant, an inferential test is required after fulfilling the prerequisite tests of normality and homogeneity.

Table 9 Descriptive Data of Students Based on Gender

| Gender | N | Min | Max | Mean | Std. Dev |
|---|---|---|---|---|---|
| Pre-test P | 16 | 2 | 23 | 10,93 | 5,27 |
| Pre-test L | 15 | 2 | 19 | 8,00 | 5,58 |
| Post-test P | 16 | 13 | 92 | 53,62 | 19,85 |
| Post-test L | 15 | 21 | 90 | 59,46 | 22,47 |
| delayed-test P | 16 | 8 | 83 | 44,43 | 20,08 |
| delayed-test L | 15 | 6 | 90 | 51,53 | 23,21 |

Before conducting inferential tests to determine the statistical significance of these differences, the assumptions of normality and homogeneity were checked. Normality test results using Shapiro-Wilk showed that all data in the pre-test, post-test, and delayed-test were normally distributed (p> 0.05). The homogeneity test using Levene also showed that the variance between groups was homogeneous (p> 0.05). Thus, the data met the requirements for the next parametric test, namely the Independent Sample t-test.

Table 10 shows the results of the Independent Sample t-test comparing male and female students' learning outcomes across three assessments: pre-test, post-test, and delayed test. The findings indicate no significant differences between the two groups at any measurement point (p > 0.05). Although descriptively female students scored higher on the pre-test, while male students outperformed in the post-test and delayed test, the mean differences were too small to reach statistical significance. Therefore, it can be concluded that gender has no significant impact on learning outcomes in either the short or long term.

Table 10 Independent Sample T-Test Results Based on Students' Gender

| Three points in time | t | df | Sig. (2-tailed) | Mean Difference |
|---|---|---|---|---|
| Pre-test | 1,507 | 29 | 0,143 | 2,93 |
| Post-test | -0,768 | 29 | 0,449 | -5,84 |
| delayed-test | -0,912 | 29 | 0,369 | -7,09 |

**Does student gender or ability level interact with worked-example effectiveness?**

Normality and homogeneity of the data for the between-subjects factor were pre-tested using Shapiro-Wilk and Levene's Test; the results showed that the assumptions of normality and homogeneity of variance were met for each group (p > 0.05). For within-subject data, the Shapiro-Wilk test showed that normality was met (p > 0.05).

Before conducting the Mixed Design ANOVA analysis, a sphericity assumption test was conducted to ensure the similarity of variance differences between measurement times (pre-test, post-test, delayed-test). This test uses Mauchly's Test of Sphericity. If the significance value is <0.05, then the assumption is not met, and the analysis continues with Greenhouse-Geisser correction; conversely, if p > 0.05, the F value in the Sphericity Assumed row was used.

Table 11 presents the results of Mauchly's Test of Sphericity, which checks whether the variances of the differences between all combinations of measurement times (pre-test, post-test, delayed-test) are equal. Since the (Sig.) value of 0.00 < 0.05, the assumption of sphericity is considered not met, so the ANOVA analysis must use the correction value (Greenhouse-Geisser) to correct the degrees of freedom.

Table 11 Test of Sphericity (Mauchly's Test Results)

| Mauchly's Test of Sphericity | | | | |
|---|---|---|---|---|
| **Within-Subjects Effect** | **Mauchly's W** | **Approx. Chi-Square** | **df** | **Sig.** |
| Time | 0,622 | 12,362 | 2 | 0,002 |

Table 12 presents the results of the Mixed Design ANOVA with the Greenhouse–Geisser correction, which was conducted to determine whether there was an interaction effect between worked-example effectiveness and gender, as well as between worked-example effectiveness and ability level, across three measurement points (pre-test, post-test, delayed-test). Based on table 12, there is no significant interaction between worked-example effectiveness and gender (sig. 0.118), nor between worked-example effectiveness and ability level (sig. 0.296). This shows that the change in student learning outcomes from pre-test to post-test to delayed-test is not influenced by differences in gender or student ability level. Based on table 12, there is no significant interaction between worked-example effectiveness and gender (sig. 0.118), nor between worked-example effectiveness and ability level (sig. 0.296). This shows that the change in student learning outcomes from pre-test to post-test to delayed-

test is not influenced by differences in gender or student ability level. The following are the profile plots of the results of the interaction of gender and student ability level.

Table 12 Mixed Design ANOVA Results Using Greenhouse-Geisser Correction

| Tests of Within-Subjects Effects | | | | | |
|---|---|---|---|---|---|
| **Source** | | **df** | **Mean Square** | **F** | **Sig.** |
| Time*Gender | Greenhouse-Geisser | 1,451 | 487,875 | 2,403 | 0,118 |
| Time*TK | Greenhouse-Geisser | 1,451 | 246,810 | 1,216 | 0,296 |

The plots in Figure 2 show similar patterns of change in learning outcomes across pre-test, post-test, and delayed-test for different genders and ability levels. This visual pattern supports the statistical results in Table 12, indicating no significant interaction effects.

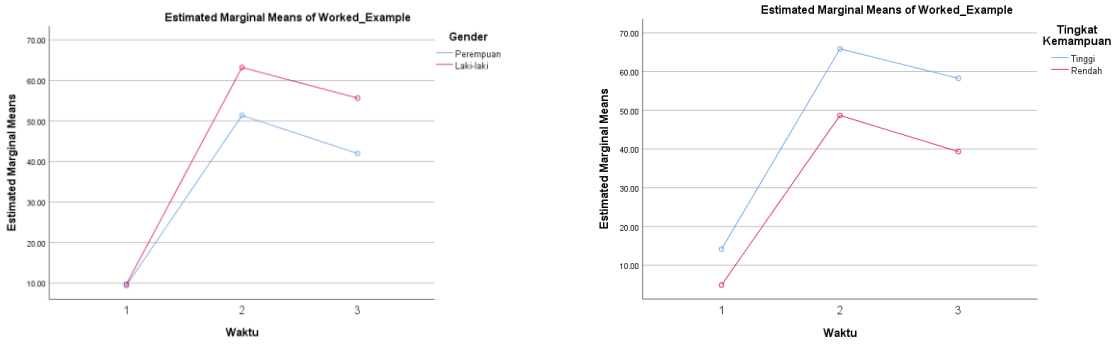


Figure 2 Profile plots of the results of the interaction of gender and student ability level

## DISCUSSION

The findings reveal that the worked-example strategy significantly improves students' learning outcomes in the short term, as shown by the increase in average scores from the pre-test to the post-test. This improvement suggests that students were able to internalize and apply the solution procedures presented in the worked examples, especially when solving similar problems. However, the slight decline in the delayed-test scores, although statistically insignificant, indicates a potential decrease in students' retention over time without reinforcement. These results support the idea that while worked examples are effective for immediate learning gains, sustained understanding requires ongoing review or practice to prevent cognitive decay.

These findings are consistent with studies by (Adeniji, 2023; Alreshidi, 2021; Chen et al., 2017) which reported a significant effect of worked examples on post-test results. In contrast (Chen, 2016) found that the effect of worked examples appeared more strongly in the delayed test than in the post-test." One possibility why the worked example strategy gave significant results in the post-test is that students managed to "borrow" the solution scheme from the worked example given, then were able to transfer it to similar problems tested in the post-test. This is in line with the principle of borrowing and reorganizing in human cognitive architecture, as explained by (Paas & van Merriënboer, 2020) When students do not yet have their own schemes, they will borrow knowledge structures from available examples.

The decrease in student learning outcomes in the delayed-test, although not too large, is likely due to the forgetting factor that arises because there is no repetition or reinforcement of the material after the treatment is given. In addition, the direct and structured worked-example strategy, although effective in the short term, tends not to challenge students to build their own understanding in solving problems. In this context, students follow the pattern provided rather than developing their own solution strategies. This relates to how students manage their cognitive resources. When they rely too much on the examples provided, the process of schema construction or deep knowledge structure can be limited.

As stated by Aurelia et al. (2024) and Sweller et al. (2019), students tend to only learn the steps of solving without really understanding the reasoning behind those steps. As a result, their understanding of the material, especially in solving complex math problems, can decline over time. Therefore, the worked example strategy should not only emphasize providing the correct solution steps, but also encourage students to understand the reasons why the steps are done and what the consequences are if they choose the wrong steps, to form a deeper and longer-lasting understanding.

On the other hand, the results showed that there was no significant difference in the worked-example effect between male and female students at all three measurement time points, either pre-test, post-test, or delayed-test. This suggests that the worked-example strategy has a relatively equal effect on both gender groups. The clear step-by-step guidance in the worked example allows all students to receive equal support during the learning process, so individual differences such as gender do not significantly affect learning outcomes. The results of this study are in line with the findings of Adeniji (2023) who stated that the application of the worked-example strategy has a relatively equal effect on male and female students. However, this result contradicts Abbott (2021) Research shows that the strategy is more effectively used by female students. This difference in findings is most likely due to the variation in the type of worked example applied. Abbott used the faded worked-example model, which is an approach that gradually reduces assistance to students. Meanwhile, in this study, students consistently received full guidance throughout the learning process. The full guidance is thought to be a factor in the absence of significant differences in learning outcomes between male and female students.

In contrast, based on the results of the analysis, there is a significant difference in learning outcomes between students with high and low ability levels during the pre-test. This is reasonable because the measurements were taken before the treatment, so the scores obtained reflect the students' initial abilities. The average achievement of high-ability students is much higher than that of low-ability students. Interestingly, however, in the post-test and delayed-test measurements, the difference between the two groups was no longer statistically significant, with both groups showing relatively equal improvements in learning achievement. This finding can be explained by the characteristics of the worked-example strategy itself, which really helps students, especially those with low abilities, to understand the material through clear and structured examples of solutions. Students are not required to find their solutions from the start, but rather are given a step-by-step guide that they can follow and learn from. This allows lower-ability students to catch up with higher-ability groups during the learning process. In other words, the worked-example strategy has the potential to reduce the gap in learning outcomes by ability level, as it provides the additional cognitive support needed by students with lower starting levels. This strategy indirectly balances learning outcomes, although higher-ability students still show slightly superior performance. This suggests that the worked-example strategy remains effective for all ability levels, but the enhancement effect is more pronounced for lower ability students.

Unlike the study (Adeniji, 2023), which found an interaction between students' skill level and the worked-example strategy, this study did not find any significant interaction between the worked-example learning strategy and students' gender or skill level, either in the short or long term. This is likely because learning gains were relatively evenly distributed across both ability groups**,** so the pattern of score changes over time was not statistically different enough to be considered an interaction**.**

Thus, these findings have important implications for learning, particularly in the use of the worked-example strategy for complex mathematics materials. However, this study also has several limitations. Only the full worked example was used, so variations such as the faded worked example were not tested. The participants were limited to a single educational level and specific material, which restricts the generalizability of the results. In addition, the study design did not include a control group or a more robust experimental setup, limiting internal validity and broader applicability. Based on these limitations, it is suggested that future research explore the comparison between different types of worked examples, such as faded worked examples and full worked examples, to find out which strategy is more effective in improving students' understanding. Further research should also involve subjects from different educational levels or materials and use a more robust experimental design, for example, by including a control group, so that the results of the study can be generalized more widely and contribute more deeply to the development of worked-example-based learning strategies

## CONCLUSION

Based on the results of data analysis from research conducted at SMP Negeri 1 Pulosari on grade VIII students in the 2024/2025 school year, the following conclusions were obtained. The application of the worked-example strategy has a positive effect on student learning outcomes in complex mathematics material at three different time points, which is shown through an increase in scores from the pre-test to the post-test, although there is a slight decrease from the post-test to the delayed-test. This strategy was found to be more effective for low-ability students as it helped them catch up and reduce the gap in learning achievement. In addition, there was no significant difference between the learning outcomes of male and female students and no significant interaction between gender or ability level with the effectiveness of the worked-example strategy. Based on these results, it is suggested that future research explore the comparison between different types of worked examples, such as faded worked examples and full worked examples, to find out which strategy is more effective in improving students' understanding. Further research should also involve subjects from different educational levels or materials and use a more robust experimental design, for example by including a control group, so that the results of the study can be generalized more widely and contribute more deeply to the development of worked-example-based learning strategies.

## ACKNOWLEDGMENTS

The authors would like to thank the supervisors for their guidance and direction during the research process. Gratitude is also extended to SMP Negeri 1 Pulosari and all staff and students who have provided opportunities and support, so that this research can run smoothly and be completed properly. In addition, the author is also grateful to all those who directly or indirectly assisted in the implementation of this research.

## DECLARATIONS

| | | |
|---|---|---|
| Author Contribution | : | Rodiyatul Hafidhoh: Conceptualization, writing - original draft, editing, and visualization;<br>Cecep Anwar Hadi Firdos Santosa: Writing - review & editing, formal analysis, and methodology;<br>Anwar Mutaqin: Validation and Supervision |
| Funding Statement | : | This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. |
| Conflict of Interest | : | The authors declare that there is no conflict of interest regarding the publication of this article. |
| AI Use Statement | : | We hereby confirm that no artificial intelligence (AI) tools or methodologies were utilized at any stage of this study, including during data collection, analysis, visualization, or manuscript preparation. All work presented in this study was conducted manually by the authors without the assistance of AI-based tools or systems. |
| Additional Information | : | Additional materials related to this study (such as test instruments, scoring rubrics, or raw data) are available upon request. |

## REFERENCES

Abbott, A. (2021). Gender differences in perceptions of the use of faded worked examples in mathematics. *Proceedings of the British Society for Research into Learning Mathematics 41(1) March 2021*, *41*(1), 1–6.

Adeniji, S. M. (2023). Effects of Worked Example on Students' Learning Outcomes in Complex Algebraic Problems. *International Journal of Instruction*, *16*(2), 229–246. https://doi.org/10.29333/iji.2023.16214a

 

Adeniji, S. M., Ameen, S. K., Dambatta, B. U., & Orilonise, R. (2018). Effect of mastery learning approach on senior school students' academic performance and retention in Circle Geometry. *International Journal of Instruction*, *11*(4), 951–962. https://doi.org/10.12973/iji.2018.11460a

Adeniji, S. M., & Baker, P. (2023). *Effects of Worked Example on Students ' Learning Outcomes in Complex Algebraic Problems*. *16*(2), 229–246. https://doi.org/10.29333/iji.2023.16214a

Adi, N. R. M., Amaruddin, H., Maulana, H., Adi, M., & Laili Qurroti A'yun, I. (2022). Validity and reliability analysis using the Rasch model to measure the quality of mathematics test items of vocational high schools. *Journal of Research and Educational Research Evaluation*, *11*(1), 103–113. http://journal.unnes.ac.id/sju/index.php/jere

Afidah, V. N. (2020). Prinsip- Prinsip Teori Beban Kognitif Dalam Merancang Media Pembelajaran Matematika. *JP2M (Jurnal Pendidikan Dan Pembelajaran Matematika)*, *1*(2), 72–79. https://doi.org/10.29100/jp2m.v1i2.195

Alreshidi, N. A. K. (2021). Effects of Example-Problem Pairs on Students' Mathematics Achievements: A Mixed-Method Study. *International Education Studies*, *14*(5), 8–18. https://doi.org/10.5539/ies.v14n5p8

Anisa, R., & Endah Retnowati. (2024). Pengaruh metode integrated worked example terhadap kemampuan pemecahan masalah dan cognitive load. *Jurnal Pengembangan Pembelajaran Matematika*, *6*(1), 14–26. https://doi.org/10.14421/jppm.2024.61.14-26

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research*, *70*(2), 181–214. https://doi.org/10.3102/00346543070002181

Aurelia, D., Ekawati, R., & Arifin, S. (2024). Primary Students' Comprehension Errors in Translating Math Problems. *Edumatica : Jurnal Pendidikan Matematika*, *14*(3), 213–230. https://doi.org/10.22437/edumatica.v14i3.38814

Ayebale, L., Habaasa, G., & Tweheyo, S. (2020). Factors affecting students' achievement in mathematics in secondary schools in developing countries: A rapid systematic. *Statistical Journal of the IAOS*, *36*(S1), S73–S76. https://doi.org/10.3233/sji-200713

Ayu, A., Sari, I., & Lutfi, A. (2024). Pengaruh Self Efficacy dan Kecemasan Matematis terhadap Hasil Belajar Mahasiswa Mata Kuliah Statistika Ekonomi The Influence of Self-Efficacy and Mathematical Anxiety on Student Learning Outcomes in Economic Statistics Courses. *Edumatica:Jurnal Pendidikan Matematika*, *14*(March).

Barbieri, C. A., Miller-Cotto, D., Clerjuste, S. N., & Chawla, K. (2023). A Meta-analysis of the Worked Examples Effect on Mathematics Performance. *Educational Psychology Review*, *35*(1), 11–33. https://doi.org/10.1007/s10648-023-09745-1

Chen, O. (2015). The worked example effect, the generation effect, and element interactivity. *Journal of Educational Psychology*, *107*(3), 689–704. https://doi.org/10.1037/edu0000018

Chen, O. (2016). Relations between the worked example and generation effects on immediate and delayed tests. *Learning and Instruction*, *45*, 20–30. https://doi.org/10.1016/j.learninstruc.2016.06.007

Chen, O., Retnowati, E., Chan, B. B. K. Y., & Kalyuga, S. (2023). The effect of worked examples on learning solution steps and knowledge transfer. *Educational Psychology*, *43*(8), 914–928. https://doi.org/10.1080/01443410.2023.2273762

Chen, Retnowati, E., & Kalyuga, S. (2017). *Effects of worked example on step performance in solving complex problems. 1*, 1–28.

Hoogerheide, V. (2014). Comparing the effects of worked examples and modeling examples on learning. *Computers in Human Behavior*, *41*, 80–91. https://doi.org/10.1016/j.chb.2014.09.013

Hyde, J. S. (2014). Gender similarities and differences. *Annual Review of Psychology*, *65*(June), 373–398. https://doi.org/10.1146/annurev-psych-010213-115057

Intan, I. N., & Rosyid, A. (2020). Peningkatan Kemampuan Pemahaman Matematis Siswa Menggunakan Worked Example. *M A T H L I N E Jurnal Matematika Dan Pendidikan Matematika*, *5*(1), 26–36. https://doi.org/10.31943/mathline.v5i1.127

Irwansyah, M. F., & Retnowati, E. (2019). Efektivitas worked example dengan strategi pengelompokan siswa ditinjau dari kemampuan pemecahan masalah dan cognitive load. *Jurnal Riset Pendidikan Matematika*, *6*(1), 62–74. https://doi.org/10.21831/jrpm.v6i1.21452

Kalyuga, S., & Renkl, A. (2010). Expertise reversal effect and its instructional implications: Introduction to the special issue. *Instructional Science*, *38*(3), 209–215. https://doi.org/10.1007/s11251-009-9102-0

Muhtarom, M. (2024). Developing an instruments to measure prospective teacher beliefs about mathematical problem-solving using the Rasch model. *Jurnal Elemen*, *10*(2), 274–288. https://doi.org/10.29408/jel.v10i2.25040

Muryanto, D. (2020). Efektivitas Worked Example Pairs Pada Pembelajaran Daerah Penyelesaian The Effectiveness of Worked Example Pairs. *Jurnal Edukasi Matematika*, 70–78.

Nainggolan, F., & Pasaribu, E. (2021). Analisis Capaian Belajar Siswa Sman Di Indonesia Tahun 2019 Dengan Pemodelan Mixed Geographically Weighted Regression. *Seminar Nasional Official Statistics*, *2020*(1), 771–780. https://doi.org/10.34123/semnasoffstat.v2020i1.509

Paas, F., & van Merriënboer, J. J. G. (2020). Cognitive-Load Theory: Methods to Manage Working Memory Load in the Learning of Complex Tasks. *Current Directions in Psychological Science*, *29*(4), 394–398. https://doi.org/10.1177/0963721420922183

Rahim, A., & Haryanto, H. (2021). Journal of Educational Research and Evaluation Implementation of Item Response Theory (IRT) Rasch Model in Quality Analysis of Final Exam Tests in Mathematics Article Info. *Journal of Educational Research and Evaluation*, *10*(2), 57–65. http://journal.unnes.ac.id/sju/index.php/jere

Ratnasari, G. (2023). Jurnal Didactical Mathematics Analisis Beban Kognitif dalam Kemampuan Pemahaman Konsep Matematis Siswa. *Jurnal Didactical Mathematics*, *5*(2), 2023. https://ejournal.unma.ac.id/index.php/dm

Retnowati, E., Ayres, P., & Sweller, J. (2010). Worked example effects in individual and group work settings. *Educational Psychology*, *30*(3), 349–367. https://doi.org/10.1080/01443411003659960

Richardo, R., & Cahdriyana, R. A. (2021). Strategi meminimalkan beban kognitif eksternal dalam pembelajaran matematika berdasarkan load cognitive theory. *Humanika*, *21*(1), 17–32. https://doi.org/10.21831/hum.v21i1.38228

Rohman, H. M. H., & Retnowati, E. (2018). How to teach geometry theorems using worked examples: A cognitive load theory perspective. *Journal of Physics: Conference Series*, *1097*(1). https://doi.org/10.1088/1742-6596/1097/1/012104

 

Santosa, C. A. H. F., & Filiz, M. (2025). Investigating the limit of peer collaboration: Insight from worked-example in multivariable calculus. *Infinity Journal*, *14*(2), 461–482. https://doi.org/10.22460/infinity.v14i2.p461-482

Santosa, C. A. H. F., Prabawanto, S., & Marethi, I. (2019). Fostering Germane Load Through Self-Explanation Prompting In Calculus Instruction. *Indonesian Journal on Learning and Advanced Education (IJOLAE)*, *1*(1), 37–47. https://doi.org/10.23917/ijolae.v1i1.7421

Santosa, C. A. H. F., Rafianti, I., & Yulistiany, D. (2022). Worked-Example Method on Mathematical Problem-Solving Ability in term of Students' Initial Ability. *Kreano, Jurnal Matematika Kreatif-Inovatif*, *13*(2), 210–220. https://doi.org/10.15294/kreano.v13i2.33301

Santosa, C. A. H. F., Suryadi, D., Prabawanto, S., & Syamsuri, S. (2018). The role of worked-example in enhancing students' self-explanation and cognitive efficiency in calculus instruction. *Jurnal Riset Pendidikan Matematika*, *5*(2), 168–180. https://doi.org/10.21831/jrpm.v0i0.19602

Setiawan, T. B., Suharto, & Susanto, A. (2016). Pengembangan Perangkat Pembelajaran Matematika Model Discovery Learning dengan Memperhatikan Beban Kognitif pada Materi Trigonometri Kelas X SMK. *Kadikma*, *7*, 1–9.

Sholihah, D. A. (2022). Strategi Pembelajaran Matematika Berdasarkan Cognitive Load Theory untuk Meminimalkan Extraneous Cognitive Load. *EQUALS: Jurnal Ilmiah Pendidikan Matematika*, *5*(1), 13–23. https://doi.org/10.46918/equals.v5i1.1197

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1016/0364-0213(88)90023-7

Sweller, J. (2011). Cognitive Load Theory. In *Psychology of Learning and Motivation - Advances in Research and Theory* (Vol. 55). Elsevier Inc. https://doi.org/10.1016/B978-0-12-387691-1.00002-8

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive Architecture and Instructional Design: 20 Years Later. *Educational Psychology Review*, *31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Yulianto, D., Juniawan, E. A., & Junaedi, Y. (2025). Gender and Feedback Effects in Digital Game-Based Learning for Primary Mathematics Education. *Edumatica: Jurnal Pendidikan Matematika*, *15*(1).