Teaching Gradient Descent

An Exploratory Study on Classic Textbook vs. Multiple Representations Approaches

Fabiana-Maria Severin

Supervisors: Gosia Migut, Ilinca Rențea

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 19, 2026

Name of the student: Fabiana-Maria Severin
Final project course: CSE3000 Research Project
Thesis committee: Gosia Migut, Ilinca Rențea, Jorge Martinez Castaneda

Abstract

Machine learning is increasingly important in computer science education, but introductory concepts can be difficult because they combine mathematical notation, algorithmic reasoning, and conceptual understanding. Gradient descent is one such concept: students may reproduce the update rule while still struggling to explain the role of the loss function, gradient, learning rate, and repeated parameter updates. This paper investigates whether multiple representations can support beginner understanding compared with a classic textbook-style explanation.

A small-scale exploratory experiment was conducted with students who had little or no prior machine learning experience. Participants completed a prerequisite pre-test, studied gradient descent using either a text-and-formula-based explanation or a multiple-representations explanation, completed a post-test, and answered an experience survey. The multiple-representations condition showed higher post-test performance, especially on computation and application tasks, as well as higher confidence, clarity, usefulness, and engagement. Perceived cognitive load remained similar across conditions. These findings suggest that aligned multiple representations can help beginners connect formal notation with concrete calculations and intuitive understanding, although the results should be interpreted cautiously because of the small sample size.

1 Introduction

Machine learning is now an important part of computer science education and is used across many real-world domains, including software engineering, healthcare, transportation, and finance.[1] As a result, students are increasingly expected to understand core machine learning concepts early in their studies. However, introductory machine learning topics can be difficult for beginners because they often require students to connect mathematical notation, algorithmic procedures, and conceptual reasoning at the same time [19].

Gradient descent is a clear example of this challenge. Although its update rule can be written compactly, understanding it requires learners to connect parameters, loss functions, gradients, learning rates, and repeated updates. A student may be able to apply the formula without being able to explain why the update moves in the negative gradient direction or why a large learning rate can cause overshooting. In this case, the student has learned a procedure, but has not yet developed usable conceptual understanding.

This project builds on prior work by Rențea, Migut, and Krijthe, who investigated interactive visualizations in machine learning education, including gradient descent [17]. The present study narrows the focus to gradient descent and compares a classic textbook-style explanation with an integrated multiple-representations explanation. In this study, representations include textual explanation, mathematical notation, visualizations, analogies, worked examples, and interactive exploration. These may help learners connect formal procedures to intuition, but they may also increase cognitive load if they are not clearly aligned [21; 11].

The study is guided by the following research questions:

RQ1 How does learning gradient descent with multiple representations affect students' conceptual understanding and problem solving performance compared with a classic textbook-style method?

RQ2 How does learning gradient descent with multiple representations affect students' perceived clarity, confidence, cognitive load, usefulness, and engagement compared with a classic textbook-style method?

To answer these questions, this study conducts a small-scale exploratory experiment with students who have little or no prior experience with machine learning. Participants complete a pre-test on prerequisite mathematical knowledge, study one of two learning materials, and then complete a gradient descent post-test and experience survey. Think-aloud notes and written reflections are also analysed to understand how students reasoned about the material.

The main contribution of this work is a mixed-methods comparison of two approaches to teaching gradient descent to beginners. The study combines post-test scores, sub-score analysis, survey constructs, qualitative thematic analysis, and intercoder reliability to examine where multiple representations appear to support learning and where their benefits are more limited. The results provide insight into how multiple representations can support introductory gradient descent learning.

The remainder of this paper is structured as follows. Section 2 discusses related work. Section 3 presents the research method. Section 4 describes the learning materials. Section 5 presents the results. Section 6 discusses interpretations. Section 7 discusses responsible research. Section 8 concludes the paper and outlines future work.

2 Background and Related Work

2.1 Gradient descent as a beginner learning challenge

Gradient descent is an optimization method used to improve model parameters by repeatedly updating them in a direction that reduces a loss function. In the one-parameter case, the update can be written as:

$$\theta_{t+1} = \theta_t - \alpha \frac{dJ}{d\theta}(\theta_t),$$

where θ_t is the current parameter value, $J(\theta)$ is the loss function, $\frac{dJ}{d\theta}$ is the derivative of the loss with respect to the parameter, and α is the learning rate. In higher dimensional settings, the derivative is replaced by the gradient vector $\nabla J(\theta_t)$. This description follows the standard presentation of gradient-based optimization in machine learning textbooks like Bishop's *Pattern Recognition and Machine Learning* [3].

[1] IBM Data and AI Team, "10 everyday machine learning use cases," https://www.ibm.com/think/topics/machine-learning-use-cases.

For an experienced learner, the update rule is compact and meaningful. For a beginner, it combines several ideas that may not yet be connected. The learner must understand that the loss value measures how bad the current model is, that the gradient describes the local direction and steepness of change in the loss, that the negative gradient direction is used because it points toward decreasing loss, and that the learning rate controls the size of each update. Learners must also understand that gradient descent is iterative: one update is usually not enough, so the process is repeated until the model reaches a sufficiently good parameter value or another stopping condition is met.

This makes gradient descent a useful concept for studying instructional design. It is mathematical enough to require formal reasoning, but also visual enough to be represented through loss curves, parameter movement, worked calculations, and analogies. Prior work has found that students can struggle with gradient descent and stochastic gradient descent, especially when the topic is presented mainly through mathematical formulation.[2] Therefore, beginner-oriented material should not only present the update rule, but also help learners connect the loss function, gradient direction, learning rate, and repeated update process.

2.2 Machine learning education

Machine learning education is still a developing area within computing education research. While machine learning is widely used in practice, research on how students learn machine learning concepts is less mature. Ko argues that computing education needs a stronger understanding of how machine learning should be taught, especially because machine learning differs from traditional rule-based programming [9]. Shapiro and Fiebrink similarly argue that machine learning education should be studied as its own research area, including for learners who need conceptual access without becoming machine learning specialists [18].

Recent computing education research also shows that students experience machine learning as challenging because it combines mathematics, programming, notation, and algorithmic reasoning. Sibia et al. report that students identify mathematical notation, vectorization, and model implementation as important sources of difficulty in machine learning courses [19]. These challenges are relevant for gradient descent because understanding it requires students to move between a formula, a computational procedure, and an interpretation of how model parameters change during training.

A recurring challenge in machine learning education is the difference between black-box and white-box understanding. A black-box understanding may allow students to use a machine learning tool, while a white-box understanding requires them to reason about what the algorithm is doing internally. Mike and Hazzan argue that learners benefit from developing a more transparent understanding of machine learning algorithms, especially when they need to interpret, debug, or improve models [12]. Gradient descent is closely connected to this issue because it is part of the internal training process of many models. If students only know that a model "learns from data" but do not understand how parameters are adjusted, their understanding remains incomplete.

2.3 Multimedia learning and multiple representations

Multiple-representations instruction is related to the broader field of multimedia learning. Mayer's cognitive theory of multimedia learning is based on the idea that learners process verbal and visual information through partly separate channels, that each channel has limited capacity, and that meaningful learning requires active selection, organization, and integration of information [11]. In this view, using words and pictures together can support learning when the representations are designed to complement each other.

For gradient descent, this means that a graph should not only decorate the explanation. It should help learners connect the formula to the visual meaning of moving along a loss curve. Similarly, an analogy should not replace the formal definition, but should help students build an intuitive bridge toward it.

Worked examples are especially relevant for beginner learners. Atkinson et al. argue that worked examples can reduce unnecessary problem-solving effort during early learning and help students focus on the structure of a procedure [2]. For gradient descent, a worked example can show how to compute the derivative, apply the update rule, and observe whether the loss decreases. This may be more accessible than asking students to solve a full optimization problem immediately after reading the formula.

2.4 Interactive visualizations for gradient descent

Visualizations of gradient descent are common in teaching. Instructors often use diagrams, lecture animations, and interactive demonstrations to show how different learning rates affect movement on a loss surface. However, many of these resources are used informally in courses, and their learning effects are not always evaluated empirically. As a result, it is less clear when they actually improve understanding.

Rențea, Migut, and Krijthe directly address this issue by studying interactive visualizations in machine learning education [17]. Their study compared interactive and static materials for two machine learning topics: gradient descent and principal component analysis. The target group was Computer Science and Engineering bachelor students who had prerequisite mathematical knowledge but had not yet followed a machine learning course. Their results showed a significant positive effect for knowledge related to principal component analysis, but no clear difference in knowledge gain for gradient descent.

This result is important for the present study because it suggests that interactive visualization alone is not guaranteed to improve gradient descent learning. One possible explanation is that gradient descent requires learners to coordinate several representations at once: the loss function, the parameter value, the derivative or gradient, the update direction, the learning rate, and the repeated nature of the algorithm. A

[2] Angelescu, "Discovering the Misconceptions that Influence Learning of Machine Learning Concepts," TU Delft Repository, https://resolver.tudelft.nl/uuid:1207ff9c-9169-45b9-9335-8ce69c8b5282.

visualization may make movement visible, but it may not be sufficient on its own if students do not also connect that movement to the formula and the underlying concepts.

2.5 Cognitive load

Cognitive load theory provides an important caution for this study. Sweller argues that instruction must consider the limits of working memory and avoid unnecessary processing that does not contribute to learning [21]. In the context of multiple representations, this means that adding more representations may help or hurt depending on their design. If learners must constantly switch between disconnected explanations, the material may increase extraneous cognitive load. If the representations are clearly aligned, they may help learners connect ideas more productively.

This concern is also relevant in computing education, where learners often need to coordinate abstract concepts, notation, procedures, and visual or code-based representations. Duran et al. review the use of cognitive load theory in computing education research and show that CLT is often used to motivate instructional design decisions, but they also warn that the theory should not be used only as a general label for mental effort [20]. They recommend considering the limits of CLT-based explanations and, where possible, measuring cognitive load rather than only speculating about it. For this study, this supports including perceived cognitive load as an additional factor when comparing the classic textbook-style condition with the multiple-representations condition.

2.6 Research gap

Overall, prior work suggests that machine learning students need support in connecting mathematical notation, algorithmic procedures, and conceptual understanding. Multimedia learning and worked-example research provide reasons to expect that multiple representations can help beginners, while cognitive load theory cautions that additional representations may become harmful if they are not clearly aligned.

In machine learning education specifically, Renţea et al. studied interactive visualizations and found that their effect for gradient descent was not clearly established [17]. Related TU Delft work by Koppelaar investigated analogies for teaching gradient descent and found no statistically significant improvement when analogies were added to concept definitions.[3] Together, these studies suggest that single instructional supports, such as visualization alone or analogy alone, may not be sufficient for improving beginner understanding of gradient descent.

The present study addresses this gap by testing a coordinated multiple-representations approach. Instead of adding one extra representation to a standard explanation, it combines visualization, analogy, worked calculation, and interaction around the same learning goals. The gap addressed by this study is therefore whether an aligned combination of representations can support beginner understanding of gradient descent better than a classic textbook-style explanation.

[3]Koppelaar, "Using analogies for teaching gradient descent," TU Delft Repository, https://resolver.tudelft.nl/uuid:c04ddaa8-abf5-4a42-8d69-833bc776c6b6.

3 Methodology

3.1 Research design

This study uses a small-scale exploratory experimental design. The goal is to compare two instructional approaches for teaching the same gradient descent learning goals. The independent variable is the teaching method, with two conditions. These conditions are summarised in Table 1.

Table 1: Overview of the two instructional conditions.

Condition	Description
Classic textbook style	Students learn gradient descent through text, formulas, and static visualizations, similar to traditional textbook or lecture-note material.
Multiple representations	Students learn the same concept through worked examples, analogies, interactive elements, and visualizations.

The dependent variables are post-test performance and student experience. Post-test performance is used to examine conceptual understanding and problem solving. Student experience is measured through survey items about clarity, confidence, cognitive load, engagement, and perceived usefulness.

The study is exploratory because the number of participants is limited and because the aim is to identify patterns rather than make broad causal claims. Nevertheless, the design is structured so that the two conditions can be compared as fairly as possible.

3.2 Participants

Participants were recruited through LinkedIn direct messages and university labs. They were first-year TU Delft computer science students with little or no prior experience with machine learning. This group was suitable because they had relevant mathematical prerequisites, such as functions, slopes, derivatives, and error measures (from courses such as Calculus or Probability and Statistics), but had not yet taken the Machine Learning course. Therefore, concepts such as loss functions, parameter optimization, gradient descent, and update variants were expected to be new or only partially familiar.

3.3 Learning goals

Both learning conditions were aligned with the same learning goals. These goals were designed to cover different levels of cognitive complexity, following Bloom's revised taxonomy [1]. In this study, the focus was mainly on understanding, applying, and analysing, because participants were expected not only to recognise basic definitions, but also to use gradient descent in simple problems and reason about its behaviour.

The learning goals covered both conceptual understanding and problem solving. They were used to guide the design of the learning materials, post-test questions, scoring rubric, and expert evaluation. A full mapping of the learning goals to Bloom's taxonomy levels is provided in Appendix A.

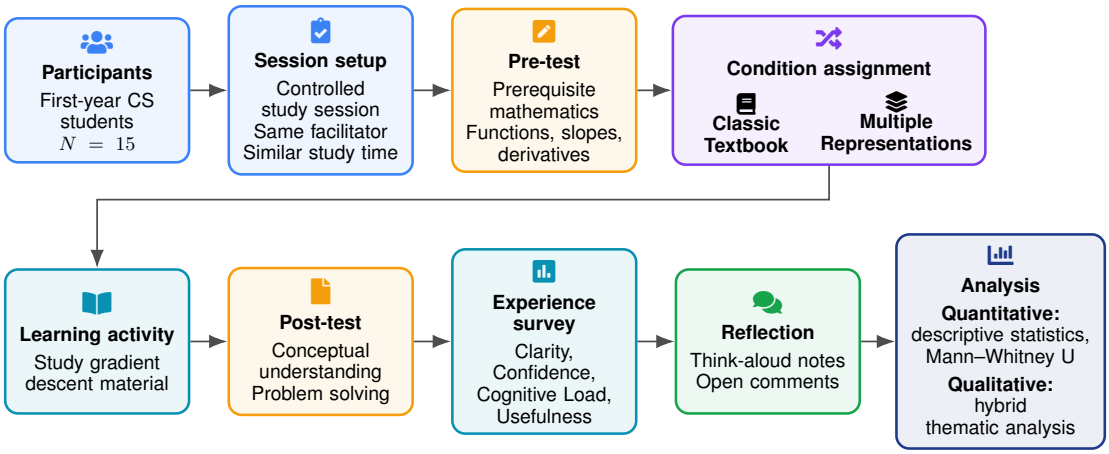

Figure 1: Overview of the experimental methodology.

3.4 Expert evaluation of learning materials and assessments

Before conducting the experiment, the learning materials and assessment questions were reviewed by a machine learning professor. The evaluation was based on selected criteria from the Quality Matters Higher Education Rubric [16]. It focused on whether the materials were accurate, suitable for beginner students, comparable across conditions, aligned with the post-test questions, and unlikely to create unnecessary cognitive load.

Based on the feedback, the final materials were revised in five ways:

- **Concepts before notation:** models, parameters, loss, gradient, and learning rate were introduced before the formal update rule.
- **Interpretable worked example:** the worked example was connected to a step-by-step process and moving visual.
- **More precise analogy:** the downhill analogy explicitly mapped the valley, position, steepness, downhill movement, and step size to loss, parameter setting, gradient, negative-gradient direction, and learning rate.
- **Assessment alignment:** the interactive activity was aligned with post-test goals, including update direction, learning-rate comparison, overshooting, starting points, and translation between representations.
- **Clearer assessment wording:** post-test questions on negative-gradient direction, overshooting, and gradient descent variants were reworded.

3.5 Procedure

The full study workflow is summarised in Figure 1. Before starting the study, participants read the participant information sheet and provided informed consent. The experiment was conducted in controlled study sessions. The learning activity was time-limited, and the same researcher facilitated all sessions to keep the procedure consistent.

The study followed five steps:

1. **Pre-test** described in Section 4.3
2. **Group assignment** Participants were assigned to the two conditions based on their pre-test performance, with the aim of forming groups with comparable prerequisite mathematical knowledge.
3. **Learning activity** Participants studied one of the two learning materials described in Sections 4.1 and 4.2
4. **Post-test** described in Section 4.4
5. **Experience survey** described in Section 4.5

3.6 Use of think aloud

Think-aloud notes were used as supporting qualitative data to capture participants' reasoning during the post-test. In think-aloud methods, participants verbalize their thoughts while completing a task, which can provide insight into their thinking process [5]. Participants first wrote individual answers to avoid copying, and then explained their reasoning after neutral questions such as "What are you thinking here?" or "Can you explain how you got this answer?" The notes were used to identify misunderstandings about gradient direction, learning rate, and update behavior, and to interpret the quantitative results.

4 Learning Materials and Assessment

The learning materials were developed for this study and informed by established introductory machine learning resources, especially Bishop's *Pattern Recognition and Machine Learning* [3] and Müller and Guido's *Introduction to Machine Learning with Python* [13]. These sources were used to keep the explanations of gradient descent, learning-rate behavior, and update variants consistent with standard machine learning terminology.

4.1 Classic textbook-style material

The classic textbook-style material was provided through a chapter of a Jupyter Book[4]. It presents gradient descent in a

[4]https://fabisev.github.io/teaching-gradient-descent-final/classic-textbook/

traditional written form. The material introduces models, parameters, loss, gradient, and learning rate, explains the goal of minimizing a loss function, presents the update rule, and describes convergence, overshooting, repeated updates, and gradient descent variants such as batch, stochastic, and mini-batch gradient descent. Although it includes a small number of simple supporting figures, it mainly relies on textual explanation, mathematical notation, and short written examples.

This condition acts as the comparison baseline because many students encounter mathematical machine learning concepts through textbook-like explanations or lecture notes.

4.2 Multiple-representations material

The multiple-representations material was provided through a second chapter of a Jupyter Book[5]. It covered the same core learning goals as the classic textbook-style material, but introduced gradient descent through geometric visualizations, a worked algebraic example, a physical analogy, and interactive exploration.

The visualizations represented gradient descent on a one-dimensional loss curve and a two-dimensional contour plot, including examples of slow, stable, and unstable learning-rate behavior. The worked example connected the update rule to hand calculation by showing how the parameter, gradient, and update size changed over repeated steps. The physical analogy explained gradient descent as a hiker moving downhill, where position represented the parameter, steepness represented the gradient, downhill movement represented the negative-gradient direction, and step size represented the learning rate. The interactive activity asked learners to predict update direction, compare learning rates, change starting points, and translate between the formula, graph, and analogy.

The important design principle was alignment. The representations were not introduced as separate explanations, but were repeatedly connected back to the same underlying ideas: parameter, loss, gradient, learning rate, update direction, repeated improvement, and update variants.

4.3 Pre-test

The pre-test was administered through a Microsoft form[6] and measured prerequisite knowledge needed for understanding gradient descent, rather than gradient descent knowledge itself. It included self-reported confidence items and multiple-choice questions on mathematical concepts such as interpreting function graphs, slopes, derivatives, loss or error values, and derivative magnitude.

The results were used to assign participants to the two experimental conditions while keeping their prerequisite mathematical backgrounds as comparable as possible.

4.4 Post-test

The post-test was provided through another Jupyter Book chapter[7] and was completed after the learning activity. The same post-test was used for both instructional conditions in order to compare participants' understanding under equivalent assessment conditions. Participants were encouraged to answer in their own words on paper, so that their reasoning process could be captured more clearly.

The post-test assessed several aspects of gradient descent understanding: core conceptual knowledge, hand computation, reasoning about the learning rate, application to a new example, explanation of the iterative optimization process, and comparison of update variants. These parts were designed to reflect both conceptual understanding and problem-solving ability. A detailed overview of the post-test structure is provided in Appendix C.

The post-test was scored using a rubric provided in Appendix D. Partial credit was used, and answers received points based on conceptual correctness, use of relevant terminology, and ability to apply the concept rather than only repeat a memorized definition.

4.5 Experience survey

The experience survey was administered through a Microsoft form.[8] It measured how participants perceived the learning activity after completing the post-test. The survey included closed-ended items and open-ended questions. The closed-ended items were grouped around perceived confidence, clarity, cognitive load, mental effort, usefulness, and engagement with the learning format.

The motivational experience items were informed by the ARCS motivation model, which describes four dimensions of learner motivation: attention, relevance, confidence, and satisfaction [8]. In this study, the survey did not attempt to measure the full ARCS model, but used it as a design reference for selected experience items.

The cognitive load items asked participants how mentally demanding or overwhelming the learning material felt. The survey also included a Paas-style 9-point mental-effort item, since subjective mental-effort ratings are commonly used in cognitive-load research [14].

The survey also included open-ended questions. Participants were asked which part of the material helped them the most, which part was most confusing, and what they would change to make the material easier to understand. These comments were used to interpret the quantitative survey scores.

5 Results

5.1 Quantitative analysis

The quantitative analysis compares the classic textbook-style condition and the multiple-representations condition using post-test scores and survey-based learner-experience measures. The post-test was analyzed as a total score out of 100 and as six sub-scores: core understanding, computation, learning-rate reasoning, application, process explanation, and gradient descent variants. Pre-test scores were used only to describe prerequisite mathematical knowledge, not to calculate learning gains.

[5] https://fabisev.github.io/teaching-gradient-descent-final/multiple-modalities/

[6] https://forms.office.com/e/1Pce1gMw6Y

[7] https://fabisev.github.io/teaching-gradient-descent-final/transfer-and-exercises/

[8] https://forms.office.com/e/R1U1EWP3zP

Because the sample size is small, the analysis is mainly descriptive. As an exploratory inferential check, the two conditions are compared using Mann–Whitney U tests [10]. Cliff's delta is reported as an effect size, where positive values indicate higher scores in the multiple-representations condition. The calculations were performed using the accompanying analysis code.[9] These tests are interpreted cautiously and are not treated as confirmatory evidence.

Survey items were grouped into construct scores for confidence, clarity, cognitive load, mental effort, perceived usefulness, and engagement. The scoring procedure is described in Appendix E.

5.2 Quantitative results

Survey results

The survey analysis included 15 responses: 7 from the classic textbook-style condition and 8 from the multiple-representations condition. Table 2 shows the construct means, mean differences, Mann–Whitney U tests, and Cliff's delta effect sizes. Figure 2 visualizes the mean scores for the five constructs measured on a 1–5 scale. Mental effort is not included because it was measured on a separate 1–9 scale.

The multiple-representations condition had higher mean scores for confidence, clarity, usefulness, and engagement. The largest differences were for engagement and usefulness. Cognitive load was almost identical across conditions, and mental effort differed only slightly.

Table 2: Survey construct scores and exploratory comparisons by instructional condition.

Construct	Classic M	Multi M	Diff.	U	p	δ
Confidence	3.61	4.31	+0.70	41.0	.139	0.46
Clarity	3.89	4.25	+0.36	38.0	.265	0.36
Cognitive load	2.57	2.59	+0.02	27.5	1.000	-0.02
Mental effort	4.71	4.88	+0.17	27.5	1.000	-0.02
Usefulness	3.43	4.17	+0.74	44.5	.061	0.59
Engagement	3.19	4.25	+1.06	48.0	.023	0.71

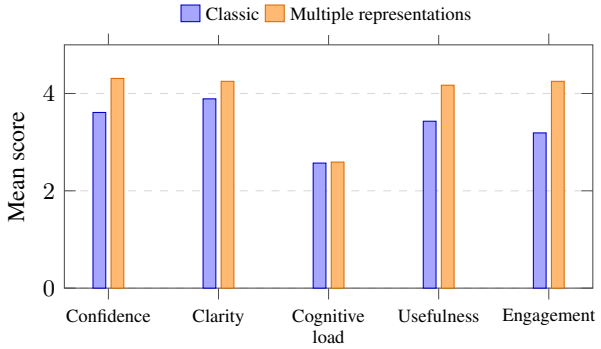

Figure 2: Mean survey construct scores by instructional condition for constructs measured on a 1–5 scale.

Participants in the multiple-representations condition were also asked which component helped them most. As shown in Table 3, the worked algebraic example was selected most often, followed by interactive exploration and physical analogy.

[9] https://github.com/fabisev/gradient-descent-modality-analysis

Table 3: Most helpful part in the multiple-representations condition.

Component	Selections
Worked algebraic example	3
Interactive exploration	2
Physical analogy	2
Geometric visualization	1

Post-test results

The post-test analysis included 15 submissions: 7 from the classic textbook-style condition and 8 from the multiple-representations condition. Table 4 reports the total score and sub-scores together with the exploratory statistical comparisons.

Table 4: Post-test scores and exploratory comparisons by instructional condition.

Measure	Classic M	Multi M	Diff.	U	p	δ
Total	66.86	83.13	+16.27	52.0	.006	0.86
A: Core	19.71	21.00	+1.29	28.5	1.000	0.02
B: Compute	9.43	18.63	+9.20	50.0	.012	0.79
C: Learning rate	11.14	12.38	+1.24	36.0	.381	0.29
D: Application	9.71	14.63	+4.92	49.5	.014	0.77
E: Process	9.71	10.75	+1.04	37.0	.288	0.32
F: Variants	7.14	5.75	-1.39	25.0	.768	-0.11

The multiple-representations condition had a higher mean total score than the classic condition. The largest differences appeared in Part B and Part D, which assessed computation by hand and application to a bowl-shaped loss function. Differences were smaller for core understanding, learning-rate reasoning, and process explanation. The only part where the classic condition had a higher mean was Part F, which focused on gradient descent variants.

5.3 Qualitative analysis

The qualitative analysis used think-aloud notes and open-ended survey responses to examine how participants reasoned about gradient descent and experienced the learning material. A hybrid thematic analysis approach was used, combining deductive codes based on the learning goals and post-test with inductive codes that emerged from participants' explanations and comments [4; 7]. This approach supports mixed-methods interpretation by connecting qualitative patterns to the quantitative results [15].

The deductive codes covered concepts such as parameter and loss, gradient direction, negative-gradient movement, learning rate, overshooting, the iterative update process, and gradient descent variants. They also captured the use of learning supports such as visualizations, analogies, worked examples, and interactive activities. Inductive codes included formula–meaning disconnect, superficial explanations, terminology difficulty, modality redundancy, need for scaffolding, and requests for broader machine learning context. The full codebook, including participant-level code frequencies, is provided in Appendix B.

To improve reliability, a subset of the data was independently coded by two coders. The two coders reached 85% agreement and a Cohen's kappa of $\kappa = 0.70$, suggesting substantial agreement. Disagreements were discussed and resolved before the remaining data were coded.

5.4 Qualitative results

The qualitative findings help explain both the survey pattern and the uneven post-test sub-score pattern.

For learner experience, the clearest qualitative pattern was the contrast between productive use of the available representations and requests for supports that were absent. In the multiple-representations condition, *Germane load*, *Visual support used*, *Interactive support used*, *Worked-example support used*, and *Analogy support used* each appeared frequently. This suggests that many participants actively used the representations to build their understanding, which is consistent with the higher perceived usefulness and engagement. In contrast, *Visual support requested* appeared for four of the seven participants in the classic condition, and *Worked-example support requested* appeared for three, suggesting that some learners wanted more concrete support than the primarily textual and mathematical material provided.

The load-related codes add nuance to this pattern. *Intrinsic load* appeared at a similar frequency in both conditions, indicating that some difficulty resulted from the inherent complexity of gradient descent rather than the presentation format. However, *Extraneous load* appeared only in the classic condition and was associated with excessive or confusing text, whereas *Cognitive overload* appeared only in the multiple-representations condition, showing that the amount of new information could occasionally become overwhelming. The codes *Analogy not useful* and *Modality redundancy* further show that individual representations were not equally useful for every learner.

The post-test sub-score pattern is also reflected in the qualitative codes. The large difference in Part B (hand computation) is consistent with *Procedural calculation fluency* and *Worked-example support used* appearing more often in the multiple-representations condition. In contrast, *Formula–meaning disconnect* and *Formula–graph mapping difficulty* appeared more often in the classic condition, suggesting that some participants could apply the formula only partially or had difficulty connecting it to the parameter, loss, or graph.

The advantage in Part D (apply the idea to a new example) is consistent with codes related to transfer and concrete representations. *Conceptual transfer difficulty* appeared more often in the classic condition, while *Visual support used*, *Interactive support used*, and *Analogy support used* appeared often in the multiple-representations condition. For Part C (reason about learning rate), *Correct learning rate trade-off* appeared in both conditions, which matches the smaller difference in learning-rate reasoning. Finally, Part F (compare update variants) showed a different pattern: the classic condition had a slightly higher mean score, and *Variant confusion* appeared more often in the multiple-representations condition.

6 Discussion

6.1 Interpretation of learner experience

The learner-experience findings suggest that the multiple-representations material made gradient descent feel more accessible and useful to beginners. The worked example, visualizations, physical analogy, and interactive exploration offered different ways to approach the same concept and may therefore explain the higher confidence, usefulness, and engagement reported in this condition.

The higher engagement may be partly related to the active role of the interactive exploration. Participants predicted outcomes, changed values, compared optimization paths, and observed changes. This may have made abstract relationships more visible and encouraged learners to check their understanding. However, engagement alone does not demonstrate learning, and some of the positive response may have resulted from the novelty of the interactive format.

Perceived cognitive load and mental effort were similar across conditions. This suggests that the additional representations were generally organized in a manageable way and may have supported productive processing by connecting formulas, calculations, graphs, and intuitive explanations [21]. At the same time, the codes *Cognitive overload*, *Analogy not useful*, and *Modality redundancy* show that the same representation was not equally useful for every learner.

6.2 Interpretation of learning outcomes

The post-test results showed a stronger overall performance pattern for the multiple-representations condition, but the differences were uneven. The clearest advantages appeared in hand computation and application to a new loss curve. Differences were smaller for core conceptual understanding, learning-rate reasoning, and explanation of the iterative process. Distinguishing between batch, stochastic, and mini-batch gradient descent was the only area in which the classic condition scored higher on average.

This pattern suggests that multiple representations were most useful when learners had to coordinate several elements of gradient descent and apply them in a concrete task. The stronger performance in hand computation can plausibly be connected to the worked algebraic example, which made the update process explicit step by step. This is consistent with worked-example research, which suggests that beginners benefit from seeing the intermediate structure of a procedure before solving a similar problem independently [2]. The higher occurrence of *Procedural calculation fluency* and *Worked-example support used* for the multiple-representations condition supports this interpretation.

The advantage in applying gradient descent to a new loss curve may be explained by the connections among the mathematical, visual, analogical, and interactive representations. Learners predicted updates, changed starting positions, interpreted movement on a loss curve, and translated between the formula, graph, and analogy. These connections may have supported a more flexible mental model that could be transferred to a new problem [11]. The lower occurrence of *Conceptual transfer difficulty* in the multiple-representations condition is consistent with this explanation.

The smaller differences in core definitions and process explanation suggest that the classic material may already have been sufficient when learners mainly needed to state ideas that were explicitly presented in both conditions. Performance was also relatively high in both groups, leaving less room for large differences. Learning-rate reasoning showed only a modest difference despite the interactive visualization

and exploration task. One explanation is that the classic material also communicated the key relationship clearly: small learning rates can lead to slow progress, while excessively large rates can cause overshooting and oscillation. The interactive visualization may therefore have strengthened recognition of these behaviours more than the ability to explain their mechanism, which is consistent with the code *What-not-why explanation*.

The result for gradient descent variants provides an important exception. These variants were represented through both the hiker analogy and the interactive comparison of smoother and noisier optimization paths. Their lower scores in the multiple-representations condition therefore do not appear to result from a lack of coverage. Instead, some learners may have struggled to translate the intuitive representations into precise terminology about using the full dataset, one training example, or a subset of examples. The classic condition's more direct textual definitions may have aligned more closely with the assessment wording. The greater occurrence of *Variant confusion* supports this interpretation.

Overall, the findings suggest that the value of a representation depends on how directly it supports the reasoning required by the task. The worked example closely matched the procedure needed for calculation, while the visual and interactive representations closely matched reasoning about movement on a loss curve. The variant representations supported intuition, but required an additional translation into formal data-level definitions. Multiple representations therefore appear most useful when the connection between the representation and the target concept is explicit.

6.3 Limitations

The main limitation is the small sample size. With 15 participants, the Mann–Whitney U tests and Cliff's delta values should be interpreted as exploratory rather than confirmatory evidence. A larger study with an a priori power analysis would be needed to determine whether the observed patterns generalize [6].

Another limitation is that participants may have differed in mathematical background, motivation, or prior exposure to optimization. The pre-test helped describe prerequisite knowledge, but it could not control for all individual differences. In addition, the post-test was completed shortly after the learning activity, so the study measured immediate understanding rather than long-term retention. The post-test and survey were also developed for this study rather than being previously validated instruments, although expert review and alignment with the learning goals support their content validity.

Finally, the multiple-representations condition combined visualizations, analogy, a worked example, and interactive exploration. The study therefore evaluates the design as a whole and cannot determine the independent contribution of each representation. The qualitative findings suggest that different learners relied on different supports, but they cannot establish whether the observed differences resulted mainly from one representation or from the connections among them.

6.4 Implications

For teaching practice, the findings suggest that representations should be selected according to the reasoning learners need to perform. A worked example can make the intermediate steps of a calculation explicit, while an interactive loss curve can connect the learning rate and gradient to visible parameter movement.

The findings also show that intuitive representations should be connected back to formal machine learning terminology. For gradient descent variants, for example, analogy and path shape may need to be paired with a concrete dataset-level comparison.

The broader implication is not that introductory material should contain as many representations as possible, but that each representation should have a clear function and prepare learners for the reasoning they are expected to perform.

7 Responsible Research

7.1 Ethical considerations

This study involves human participants and was therefore prepared as part of the TU Delft Human Research Ethics Committee (HREC) process. The HREC checklist and informed consent materials were used to identify possible risks, define mitigation measures, and ensure that participants received clear information before taking part in the study.

Before participation, students receive an information sheet explaining the purpose of the study, the procedure, the expected duration, the type of data collected, and how the data will be used. Participants are informed that the study investigates how different learning representations, such as text, formulas, visualizations, analogies, and interactive elements, affect students' understanding of gradient descent. They are also informed that participation involves interacting with learning materials, completing short learning tasks, completing an assessment, and answering a short survey about their experience.

Participation is voluntary. Participants are told that they may decline to participate, skip questions, or withdraw from the study at any time without giving a reason. They are also informed that participation, non-participation, or withdrawal has no impact on their academic evaluation, grades, or relationship with TU Delft.

The study is designed to involve minimal risk. The main possible risks are mild cognitive effort or frustration when working with the learning materials and assessment questions, perceived pressure to perform well, and a small risk of indirect identification due to the limited sample size. To reduce these risks, participants are told that the study evaluates the learning materials rather than their intelligence or academic ability. Scores are used only for research analysis and are not connected to grades.

Another ethical concern is peer influence during sessions with multiple participants. To reduce this risk, participants answer individually in writing before any verbal explanation takes place. When think-aloud explanations are collected, the researcher uses neutral prompts and avoids giving hints about correctness. This helps preserve the validity of the data and avoids turning the session into a guided tutoring activity.

7.2 Privacy and data protection

The study collects only data needed to answer the research questions. This includes pre-test answers, post-test answers, survey responses, condition assignment, and qualitative notes from think-aloud explanations. Limited personal data, such as names or email addresses, may be collected only for administrative purposes, such as scheduling or consent documentation.

Research data is stored using participant codes. Directly identifying information is stored separately from research data and is not used in the analysis. The data is anonymised or pseudonymised before analysis where possible. Results are reported in aggregate form, and individual participants are not identifiable in the paper. This is especially important because the study has a small sample size, which creates a possible risk of indirect identification.

Survey tools approved by TU Delft (Microsoft Forms) are used for collecting responses. Participants are informed about the use of online tools and the minimal risk of data breach associated with online data collection. Data exported for analysis is stored securely on TU Delft approved systems and is accessible only to the researcher and supervising team.

Personal data collected for administrative purposes is deleted after the completion of the study. Anonymised research data may be retained for academic purposes, such as verification, teaching, or future research, in accordance with TU Delft data management guidelines.

7.3 Reproducibility

Reproducibility was supported by documenting the main components of the study design: the learning materials, pre-test, post-test, scoring rubric, survey items, experimental procedure, and analysis procedure. The learning materials were made available through public Jupyter Book pages, and the post-test structure, scoring rubric, qualitative codebook, and survey scoring procedure are included in the appendices. This allows another researcher to understand how the study was conducted and to repeat or adapt the experiment with a similar participant group.

The quantitative analysis was also made more transparent by reporting the descriptive statistics, Mann–Whitney U tests, Cliff's delta effect sizes, and the accompanying analysis code. This supports reproducibility by making the calculation procedure inspectable rather than only reporting final values.

Several design decisions that affect interpretation were made explicit. For example, the pre-test measured prerequisite mathematical knowledge rather than gradient descent knowledge directly. Therefore, the analysis did not treat the difference between pre-test and post-test scores as a direct learning-gain measure. Similarly, because the sample size was small, the statistical tests were interpreted as exploratory rather than confirmatory. These decisions are important for reproducibility because they clarify not only how the data was processed, but also how the measures and results should be interpreted.

7.4 Research integrity

Research integrity was supported by reporting the findings transparently, including results that did not fully support the expected benefit of the multiple-representations condition. This includes non-significant results, small differences, similar cognitive-load scores across conditions, and the unexpected finding that the classic textbook-style condition performed slightly better on the gradient descent variants part of the post-test.

Because the study is small-scale and exploratory, the Mann–Whitney U tests and Cliff's delta values are interpreted as exploratory comparisons rather than confirmatory evidence. The paper therefore avoids claiming that multiple representations are generally superior, and instead describes where they appeared most useful in this sample: computation, application, confidence, usefulness, and engagement.

8 Conclusions and Future Work

This paper investigated whether a multiple-representations explanation can support beginner students in learning gradient descent compared with a classic textbook-style explanation. The study compared two learning conditions with the same learning goals: one using a classic text-based explanation and one combining text, visualizations, analogy, worked examples, and interactive exploration.

Within this exploratory sample, the multiple-representations condition was associated with stronger post-test performance. Participants in this condition achieved higher overall post-test scores, with the clearest differences appearing in computation and application tasks. This suggests that multiple aligned representations may help beginners connect the gradient descent update rule to concrete calculations and new problem settings. However, the difference was not equally strong across all parts of the post-test, which suggests that the approach may be especially useful for tasks that require learners to move between formulas, examples, and interpretation.

The survey results also showed higher confidence, clarity, usefulness, and engagement in the multiple-representations condition. At the same time, perceived cognitive load was almost identical across conditions, suggesting that the additional representations did not clearly make the material feel more overwhelming. The qualitative findings supported this interpretation, as participants in the multiple-representations condition often referred to visualizations, analogies, interactive elements, or worked examples when explaining their reasoning. Together, these findings suggest that multiple representations can support both learning performance and learner experience when they are aligned around the same core concepts. However, the study had a small sample size. Therefore, the results should be seen as exploratory rather than confirmatory.

Future work could extend this study with a larger participant sample and delayed post-tests to measure retention. It would also be useful to compare individual combinations of representations separately, such as text plus visualization, text plus analogy, text plus worked example, and full multiple-representations explanations. This would help identify which representations are most useful for different parts of learning gradient descent.

References

[1] Lorin W. Anderson and David R. Krathwohl, editors. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman, New York, 2001.

[2] Robert K. Atkinson, Sharon J. Derry, Alexander Renkl, and Donald Wortham. Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research*, 70(2):181–214, 2000.

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[4] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.

[5] Elizabeth Charters. The use of think-aloud methods in qualitative research. *Brock Education Journal*, 12(2), 2003.

[6] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 2 edition, 1988.

[7] Jennifer Fereday and Eimear Muir-Cochrane. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods*, 5(1):80–92, 2006.

[8] John M. Keller. Development and use of the ARCS model of instructional design. In Charles M. Reigeluth, editor, *Instructional Design Theories in Action: Lessons Illustrating Selected Theories and Models*, pages 289–320. Lawrence Erlbaum Associates, Hillsdale, NJ, 1987.

[9] Amy J. Ko. We need to learn how to teach machine learning. https://medium.com/bits-and-behavior/we-need-to-learn-how-to-teach-machine-learning-acc78bac3ff8, August 2017.

[10] Henry B. Mann and Donald R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947.

[11] Richard E. Mayer. Cognitive theory of multimedia learning. In *The Cambridge Handbook of Multimedia Learning*. Cambridge University Press, 2005.

[12] Koby Mike and Orit Hazzan. Machine learning for non-majors: A white box approach. *Statistics Education Research Journal*, 2022.

[13] Andreas C. Müller and Sarah Guido. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media, Sebastopol, CA, 2016.

[14] Fred G. W. C. Paas, Jeroen J. G. Van Merriënboer, and Jos J. Adam. Measurement of cognitive load in instructional research. *Perceptual and Motor Skills*, 79(1):422, 1994.

[15] Kevin Proudfoot. Inductive/deductive hybrid thematic analysis in mixed methods research. *Journal of Mixed Methods Research*, 17(3):308–326, 2023.

[16] Quality Matters. Higher education rubric, seventh edition: Specific review standards, 2023.

[17] Ilinca Rențea, Gosia Migut, and Jesse Krijthe. Are interactive visualizations in machine learning education helping students? In *ITiCSE 2025: Proceedings of the 30th ACM Conference on Innovation and Technology in Computer Science Education V. 1*. Association for Computing Machinery, 2025.

[18] R. Benjamin Shapiro and Rebecca Fiebrink. Introduction to the special section: Launching an agenda for research on learning machine learning. *ACM Trans. Comput. Educ.*, 19(4), October 2019.

[19] Naaz Sibia, Amber Richardson, Alice Gao, Andrew Petersen, and Lisa Zhang. Student perspectives on the challenges in machine learning. In *Proceedings of the 30th ACM Conference on Innovation and Technology in Computer Science Education*, pages 9–15. Association for Computing Machinery, 2025.

[20] Rodrigo Silva Duran, Albina Zavgorodniaia, and Juha Sorva. Cognitive load theory in computing education research: A review. *ACM Transactions on Computing Education*, 22(4):1–27, 2022.

[21] John Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285, 1988.

Appendices

A Learning goals and Bloom's taxonomy mapping

Table 5 presents the full mapping between the learning goals used in the study, their corresponding Bloom's taxonomy levels, and their role in the research design. The learning goals were identical for both learning conditions and were used to guide the design of the learning materials, post-test questions, scoring rubric, and expert evaluation.

Table 5: Learning goals and corresponding Bloom's taxonomy levels.

Learning goal	Bloom's level	Role in the study
Explain the difference between a model parameter and a loss value.	Understand	Conceptual understanding
Explain what information the gradient gives at the current parameter value.	Understand	Conceptual understanding
Describe the role of the learning rate.	Understand	Conceptual understanding
Compute simple gradient descent updates by hand for a one-dimensional quadratic loss.	Apply	Problem solving
Reason about what happens when the learning rate is too large or too small.	Analyse	Conceptual understanding and problem solving
Describe gradient descent as an iterative process.	Understand	Conceptual understanding
Distinguish between batch, stochastic, and mini-batch gradient descent at a beginner level.	Analyse	Conceptual understanding

B Qualitative Codebook

Table 6 shows the codebook used for the qualitative analysis. Participant codes beginning with 1 refer to the classic textbook-style condition, while participant codes beginning with 2 refer to the multiple-representations condition. The frequency columns indicate how many participants in each condition were assigned each code at least once. These counts are used descriptively to show recurring patterns in the qualitative data, not as statistical evidence.

Table 6: Codebook used for the hybrid thematic analysis, with descriptive participant-level code frequencies by condition.

Code	Type	Classic n	Multi n	Meaning / when to use	Example evidence from notes
Procedural calculation fluency	Deductive	2	4	Student can follow formula steps correctly, especially in Part B.	Student 1.2 found Part B easier because it was formula-based. Student 2.4 used the worked example for Part B, and Student 2.8 calculated Part B correctly and explained the steps.
Conceptual transfer difficulty	Deductive	4	1	Student struggles to apply gradient descent to a new situation, especially Part D.	Student 1.1 forgot Part D and did not know how to apply the idea to a new example. Student 1.4 struggled with the more complex explanation-based exercises.
Formula–meaning disconnect	Inductive	3	1	Student can use the formula but does not understand what the terms mean conceptually.	Student 1.1 struggled with connecting the formula to the actual meaning of the parameters. Student 1.3 had difficulty connecting the formula to the theory. Student 2.7 said: "I can follow the formula if I plug in the numbers, but I don't fully understand what each number means."
Formula–graph mapping difficulty	Inductive	2	1	Student struggles to connect the equation or update rule to the graph.	Student 1.5 said that "how the formula relates to the graph" and what the graph represents were confusing.
Correct learning rate trade-off	Deductive	3	4	Student understands that a learning rate that is too high can overshoot, while a learning rate that is too low can make progress slow.	Student 1.2 explained that a learning rate that is too large can cause updates to jump from side to side, while a very small learning rate makes progress slow. Students 2.1, 2.3, 2.6, and 2.8 also connected a high learning rate to overshooting.
What-not-why explanation	Inductive	2	2	Student can say what happens but cannot fully explain the causal mechanism.	Student 1.5 could explain parts of C, D, and E at the level of what happens, but struggled to explain why. Student 2.6 explained overshooting visually as jumping over the bottom, but the explanation was mostly descriptive.

Continued on next page

Code	Type	Classic n	Multi n	Meaning / when to use	Example evidence from notes
Direction/sign confusion	Deductive	1	1	Student struggles with positive or negative slope, gradient sign, or movement direction.	Student 2.8 said: "The gradient gives direction, but I get confused about which way is positive or negative." They also reported struggling with directions.
Gradient as direction understood	Deductive	2	5	Student correctly explains the gradient as direction, slope, or steepest change.	Student 1.7 said that the gradient gives the direction of steepest increase, so the negative gradient gives the direction to reduce the loss. Student 2.2 described the gradient as the direction of steepest change, and Student 2.5 used the downhill idea to explain movement in the negative-gradient direction.
Superficial explanation	Inductive	4	2	Student answers correctly but without much conceptual depth.	Student 2.1 understood the basic role of parameter and loss, but the explanation was brief.
Terminology difficulty	Inductive	3	2	Student understands the idea but cannot use precise mathematical or machine learning terms.	Student 1.3 used informal wording such as going "really fast" when explaining Part D. Student 1.4 struggled with terminology when explaining the full process.
Formal terminology use	Inductive	2	3	Student uses precise mathematical or machine learning vocabulary correctly.	Student 1.7 used more formal wording, describing the parameter as the value the model optimizes and the loss as the function value to minimize. Student 2.2 described the gradient as the direction of steepest change, and Student 2.7 used the right terminology for most theory-based questions.
Variant confusion	Deductive	2	5	Student struggles with batch, stochastic, or mini-batch gradient descent.	Student 1.3 understood batch and stochastic partly, but did not know mini-batch. Student 1.7 confused stochastic gradient descent with choosing a random direction. Students 2.1, 2.2, 2.5, 2.6, and 2.7 also showed uncertainty about gradient descent variants.
Big-picture missing	Inductive	2	1	Student wants to understand how gradient descent fits into machine learning models more generally.	Student 1.4 asked for a specific example of the whole modelling process and how gradient descent adjusts parameters. Student 1.5 asked for a larger scenario showing how the formula links with the whole topic. Student 2.2 suggested explaining how gradient descent fits into the wider machine learning topic.
Visual support used	Deductive	0	5	Student in the multiple-representations condition uses visuals to answer or reason.	Student 2.2 said the interactive graph helped. Student 2.3 said interactive visuals helped them understand overshooting. Student 2.4 mostly used the interactive graphs and formula. Student 2.6 used graphics to connect positive slope with movement to the left in Part D.
Visual support requested	Inductive	4	0	Student in the textbook condition asks for visuals, graphs, or images.	Student 1.1 asked for more examples, visuals, and interactive materials. Student 1.2 asked for more visuals and graphs. Student 1.5 asked for a graph that maps to the formula. Student 1.6 suggested including images.
Interactive support used	Deductive	0	4	Student uses the interactive exploration to understand learning rate, gradient, or overshooting.	Student 2.2 connected the gradient to the interactive exploration model. Student 2.3 said interactive visuals helped them understand overshooting. Student 2.4 said they mostly used the interactive graphs and formula.
Interactive support requested	Inductive	1	0	Student in the textbook condition asks for interactive material.	Student 1.1 suggested adding more interactive materials to help learning.
Worked-example support used	Deductive	0	4	Student says the worked example helped with calculation or Part B.	Student 2.4 said the worked example was useful for understanding and for solving Part B.
Worked-example support requested	Inductive	3	0	Student asks for more solved examples, calculations, or full-process examples.	Student 1.3 asked for more worked-out examples with calculations. Student 1.4 suggested a specific example of the whole modelling process. Student 1.5 asked for a larger example that goes through a scenario.
Analogy support used	Deductive	0	3	Student uses the physical or hiker analogy to understand negative gradient, learning rate, or overshooting.	Student 2.1 said the analogy helped them understand overshooting. Student 2.5 used the downhill idea to explain why we move in the negative-gradient direction. Student 2.8 said the analogy helped because large steps can jump too far.
Analogy not useful	Inductive	0	2	Student reports not using or not benefiting from the analogy.	Student 2.4 said the analogy did not help much and mostly used the interactive graphs and formula. Student 2.6 also said they did not use the analogy much.

Continued on next page

Code	Type	Classic n	Multi n	Meaning / when to use	Example evidence from notes
Modality redundancy	Inductive	0	1	Student feels visuals or multiple representations were unnecessary for some tasks.	Student 2.3 suggested that visuals were redundant for formula-based exercises and that Part B could be solved directly using the update rule.
Intrinsic load	Deductive	3	4	Difficulty caused by the inherent complexity of the concept itself, such as algebra, slope, gradients, or update variants.	Student 1.3 found the stochastic and batch gradient descent terms confusing. Student 1.6 found the gradient formula confusing. Student 1.7 found the different gradient descent types confusing. Student 2.7 found the algebraic example most confusing, and Student 2.8 struggled with slope and direction.
Extraneous load	Deductive	2	0	Difficulty caused by how the material is presented, such as too much text or unnecessary information.	Student 1.1 said the learning-rate section had too much unnecessary information. Student 1.2 said the bulk text was confusing.
Germane load	Deductive	0	5	Productive mental effort that helps learning, such as using an analogy, graph, worked example, or interactive activity to build understanding.	Student 2.1 used the analogy to understand overshooting. Student 2.2 used the interactive exploration model. Student 2.4 used the worked example for Part B. Student 2.6 used graphics to reason about slope and movement. Student 2.8 used the analogy to reason about overshooting.
Cognitive overload	Deductive	0	3	Student seems overwhelmed by too much information or too many representations.	Student 2.3 found the geometric visualization confusing because it had the most new information. Student 2.5 found the two-parameter visualization confusing.
Need for scaffolding	Inductive	2	2	Student needs concepts ordered more gradually, for example from intuition to formula or from a full example to abstract notation.	Student 1.4 suggested a specific example of the whole modelling process. Student 1.5 asked for a graph that maps to the formula and a larger scenario. Student 2.8 suggested introducing the analogy first, then moving to harder concepts and formulas. Student 2.7 suggested that the algebraic example might have been easier after the physical analogy.

C Post-test structure

The post-test contained six parts, each targeting a different aspect of gradient descent understanding. The same post-test was used for both instructional conditions.

Part A: Core understanding. Participants explained the difference between a model parameter and a loss value, described what information the gradient provides at the current parameter setting, explained why gradient descent moves in the negative gradient direction, and stated the role of the learning rate.

Part B: Computation by hand. Participants worked with the quadratic function $J(\theta) = (\theta - 4)^2$. They computed the derivative, calculated two gradient descent updates starting from $\theta_0 = 0$ with $\alpha = 0.25$, stated whether the loss was decreasing, and then repeated the exercise with $\alpha = 1.2$ to compare what changed.

Part C: Learning rate reasoning. Participants reasoned about a situation where the updates repeatedly jump from one side of the minimum to the other. They explained what this suggests about the learning rate, why this can happen even when the update direction is chosen correctly, and whether choosing a smaller learning rate always makes progress faster.

Part D: Application to a new example. Participants applied the idea to a model with one parameter on the right side of a bowl-shaped loss curve. They reasoned about which direction the parameter should move when the slope is positive, what a large slope magnitude suggests about the update size before the learning rate is applied, what should happen to updates near the bottom, and why this behaviour makes sense for optimization.

Part E: Process explanation. Participants described gradient descent as a step-by-step process for improving a parameter value. Their explanation was expected to mention the current parameter value, the loss, the gradient, the learning rate, and why the process is repeated.

Part F: Variant comparison. Participants compared batch gradient descent and stochastic gradient descent, explained why mini-batch gradient descent can be a useful compromise, and identified which variant is most similar to a noisy optimization path that still trends downward.

D Post-test scoring rubric

The post-test was scored using a 100-point rubric. Table 7 shows how the points were distributed across the different parts of the test.

Table 7: Detailed point allocation for the post-test.

Question / criterion	**Points**
Part A: Core understanding, 24 points	
Difference between a model parameter and a loss value	6
Information provided by the gradient at the current parameter setting	6
Reason for moving in the negative gradient direction	6
Role of the learning rate	6
Part B: Compute by hand, 22 points	
Correct derivative of $J(\theta) = (\theta - 4)^2$	4
Correct use of the update rule for $\alpha = 0.25$	4
Correct computation of θ_1 and θ_2 for $\alpha = 0.25$	6
Correct statement about whether the loss decreases	3
Correct repetition and interpretation for $\alpha = 1.2$	5
Part C: Reason about learning rate, 16 points	
Identifies that repeated jumping suggests a learning rate that is too large	5
Explains why overshooting can happen even when the movement direction is chosen correctly	6
Explains that a smaller learning rate does not always make progress faster	5
Part D: Apply the idea to a new example, 16 points	
Correctly states that the parameter should move left when the slope is positive	4
Connects large slope magnitude to a larger unscaled update	4
Explains that updates become smaller near the minimum as the slope decreases	4
Explains why this behaviour makes sense for optimization	4
Part E: Explain the process, 12 points	
Mentions the current parameter value and the loss	3
Mentions computing or using the gradient	3
Mentions scaling the update by the learning rate	3
Explains that the process is repeated to reduce the loss	3
Part F: Compare update variants, 10 points	
Correctly distinguishes batch gradient descent from stochastic gradient descent	4
Explains why mini-batch gradient descent is a useful compromise	4
Identifies noisy but downward-trending optimization as resembling stochastic or mini-batch gradient descent	2
Total	**100**

E Survey scoring procedure

The survey was used to describe participants' learning experience after completing the learning material. The constructs included confidence, clarity, cognitive load, mental effort, usefulness, and engagement.

Survey items were first converted into numerical values. Most items used five-point response scales, where higher scores indicated a stronger presence of the measured construct. Table 8 shows how the response options were coded.

Table 8: Numerical coding of survey response options.

Item type	**Response options**	**Numerical coding**
Confidence items	Not confident – Very confident	1 – 5
Agreement items	Strongly disagree – Strongly agree	1 – 5
Usefulness items	Not useful – Very useful	1 – 5
Mental effort item	Very, very low mental effort – Very, very high mental effort	1 – 9

Construct scores were calculated as the mean of the items belonging to each construct. Table 9 summarizes how each construct was defined.

Table 9: Survey constructs and included items.

Construct	**Included items**
Confidence	Confidence in explaining the purpose of gradient descent, interpreting the update rule, reasoning about the learning rate, and applying gradient descent to a new example.
Clarity	Perceived clarity, logical organization, ease of following the explanation, terminology, and helpfulness of examples.
Cognitive load	Perceived difficulty, overload, amount of information, rereading, and focus. The focus item was reverse-coded so that higher scores consistently indicate higher perceived cognitive load.
Mental effort	Single nine-point mental effort item. This item was kept separate because it used a different scale from the other survey items.
Usefulness	Whether the material helped participants understand the main idea, check their understanding, and apply the concept to a new example.
Engagement	Whether the format kept participants' attention, made the topic feel approachable, and would be useful for learning other Machine Learning concepts.