

From Visualization to Adaptive Learning in Algorithms and Data Structures

Hadas CHASSIDIM* [0000-0002-3575-9602], Noam BAR,
Irina RABAEV* [0000-0002-8542-8342]

Department of Software Engineering, Shamoon College of Engineering, Be'er Sheva, Israel
e-mail: noamba2@ac.sce.ac.il, hadasch@ac.sce.ac.il, irinar@ac.sce.ac.il

Abstract. Computer science students are required to master data structures and algorithms, yet learning these topics often presents significant cognitive challenges. This study investigates whether a bundled package of adaptive personalization is associated with learning outcomes, perceived workload, and usability, using an existing educational visualization tool called VZOU (Visualization Zone of Understanding). The enhanced adaptive version of VZOU incorporates personalized learning paths structured around a visible learning progression map, enabling learners to monitor their progress, identify gaps, and take ownership of their learning path. Basic gamification features, Bloom-aligned progressive sequencing, and topic-specific formative quizzes are designed to support self-regulated learning through immediate feedback and motivational reinforcement. In a between-subjects study (N = 55), we compared control, non-adaptive, and adaptive groups. The adaptive group scored significantly higher on the post-test than the non-adaptive and control groups (mean 89.3 vs. 73.3 and 71.0 out of 100, respectively), and showed numerically but not statistically higher usability scores (SUS). Perceived mental workload showed a mixed pattern: the overall six-subscale NASA-TLX composite did not differ significantly by condition, though Mental Demand specifically was significantly higher in the adaptive group than in the control group, and a workload composite excluding self-rated success also differed significantly across conditions. The bundled adaptive VZOU package, combining a diagnostic-driven graph, gated progression, formative quizzes, and gamification, was associated with higher post-test performance, while raising certain aspects of perceived workload. As these components were not manipulated independently, the estimate reflects the package as a whole, not any single element.

Key words: Adaptive learning; Algorithms; Data structures; Workload; Visualization; Usability.

1. Introduction

Data structures and algorithms (DS&A) form the foundation of modern computer science and software engineering, supporting fields such as data analysis, cybersecurity, and graphics. Mastery of these topics is essential for problem-solving, system design, and optimization tasks in various applications. However, despite their importance, students often face considerable challenges when learning data structures and algorithms due to the

*Corresponding author.

abstract nature of the concepts, the complexity of algorithmic logic, and the subjective cognitive load involved in tracking multiple procedural steps.

Previous research has highlighted that traditional teaching methods, such as static slides or textbook examples, may not effectively support deep understanding, especially when visualizing dynamic changes in data structures during algorithm execution (Hundhausen et al., 2002). Cognitive Load Theory (CLT) further suggests that minimizing unnecessary mental burden can improve learning outcomes, particularly for complex and multi-stage tasks (Sweller, 1988).

Cognitive Load Theory distinguishes between three types of mental effort in learning: intrinsic load, extraneous load, and germane load. Intrinsic load refers to the inherent complexity of the topic, e.g., rotations in the AVL tree. Extraneous load is related to information presentation and distractions, e.g., irrelevant images, distracting sounds or fonts that are difficult to read. Germane load is a work one puts into processing new information and integrating it into long-term memory. Germane load is a ‘positive’ load that contributes directly to learning outcomes.

Recent developments in CLT emphasize the role of instructional design and educational technology in optimizing intrinsic and germane load while minimizing extraneous load, especially in digital learning environments (Sweller, 2020). This perspective underscores the importance of designing adaptive interfaces that align cognitive demands with learners’ expertise to promote effective and sustainable learning.

Although visualization tools have shown potential in supporting learners’ understanding of algorithmic processes, prior studies have emphasized that visualization alone may not guarantee deep cognitive engagement. Studies such as Hundhausen et al. (2002) found that students can observe animations passively, without engaging in the active construction of mental models required for meaningful learning. From a perspective of CLT, effective learning requires not only reducing extraneous load, but also fostering germane cognitive load, where learners invest mental effort in building and refining schemas (Paas et al., 2003). Studies on educational gamification further argue that motivation and engagement mechanisms are essential to transform passive observation into active processing (Dicheva et al., 2015; Ortiz-Rojas et al., 2025) and higher. However, the combined impact of adaptive visualization, cognitive load, and motivational design in the context of algorithm learning remains underexplored. This gap is also reflected in recent reviews of adaptive learning in computer science education, which highlight limited empirical work integrating adaptivity with engagement mechanisms such as gamification (Barbosa et al., 2024).

In response to these challenges, educational approaches that promote self-learning and integrate visualization, interaction, and personalization have emerged as promising directions. Instead of merely presenting visual content, adaptive learning environments aim to tailor the learning experience to the individual, adjusting complexity and providing motivational elements to sustain engagement (Zimmerman, 2008).

This study uses VZOU (Visualization Zone of Understanding) not simply as a technological tool, but as a research environment to examine how personalization and gamification features influence learning outcomes, cognitive load, and user engagement in the

context of data structures and algorithms. We hypothesize that learners using the adaptive version of VZOU, evaluated as a single bundled intervention rather than as separable components, will achieve higher post-test scores, and will report different perceived workload, compared to those using the non-adaptive or control conditions. To test this hypothesis, a comparative study was conducted across three learning conditions within algorithms education.

2. Background

DS&A concepts are among the most challenging topics for computer science and software engineering students. The abstract nature of algorithmic logic, the complexity of dynamic changes in data structures during execution, and the need to follow multi-step processes can overwhelm learners (Hundhausen et al., 2002). This difficulty is well recognized: “For many students, however, it is not easy to master them due to their diversity and inherent complexity” (Vateha and Šimonák, 2025). According to Cognitive Load Theory (Sweller, 1988), excessive mental effort during complex tasks can hinder effective learning. Students often struggle to track sequential steps and understand the evolving structure of the data, leading to frustration and decreased motivation (Paas et al., 2003).

Recent developments in CLT further emphasize the importance of instructional design and educational technology in optimizing intrinsic and germane load while minimizing extraneous load, particularly in digital learning environments (Sweller, 2020). This theoretical perspective supports the design of adaptive interfaces that align cognitive demands with learners’ expertise and learning goals.

According to the CLT model (Sweller, 1988, 2020), total cognitive load consists of three components: intrinsic load, related to the inherent complexity of the material; extraneous load, generated by suboptimal instructional design; and germane load, reflecting mental effort devoted to constructing schemas. Effective instructional environments aim to reduce extraneous load while promoting germane load through well-structured, adaptive interfaces.

These persistent challenges indicate the need for instructional approaches that not only clarify algorithmic processes, but also support cognitive effort and sustain learner motivation—issues that arise during self-regulated learning that are further explored in the following sections.

2.1. *Self-Regulated and Adaptive Learning*

Self-regulated learning (SRL) is a widely used concept in the field of educational technology, which refers to the degree that learners actively participate in their own learning process; learners who actively monitor their own understanding and adjust their strategies accordingly are far more likely to succeed. Students who engage in SRL are not passive recipients of instruction, but rather active agents who take deliberate control over their cognitive and motivational processes in pursuit of their learning goals (Zimmerman, 2008). Adaptive learning technology supports self-regulated learning by providing timely,

personalized feedback that learners need to monitor their own progress and adjust their strategies accordingly. This need is reinforced by recent syntheses of adaptive learning research in computer science education, which shows that while personalization techniques are increasingly common, many systems still struggle to support sustained self-regulation and learner engagement in complex domains (Barbosa et al., 2024). The development of an adaptive learning tool called *studybuddy* was designed to support SRL in upper secondary schools through a co-design process involving both students and teachers (Mejeh and Sarbach, 2025). The study finds that for the tool to effectively promote SRL, it must adapt to each student's personal needs, including individual schedules, learning pace, and emotional states. The findings underline that self-regulated learning requires learners to be able to monitor their own progress, receive timely and personalized feedback, and maintain control over when and how they receive support.

2.2. *Personalization in Learning*

Personalization is increasingly recognized as a critical factor for successful learning, particularly in technology-driven education. Research has shown that adapting content to a learner's knowledge level and individual preferences significantly improves comprehension and motivation. For example, Shute and Towle (2003) emphasized that adaptive learning systems, which tailor educational content to each learner's unique profile, contribute to enhanced academic achievement and increased intrinsic motivation.

One of the central methods for implementing effective personalization is Bloom's Taxonomy (Anderson and Krathwohl, 2001). This framework organizes learning objectives according to cognitive complexity, ranging from basic understanding to higher-order thinking skills. Leveraging Bloom's Taxonomy is one proposed approach for adaptive learning systems to identify a learner's current cognitive level and adjust content and activities accordingly. Sajja et al. (2025) proposed a prototype system that classifies student queries across six hierarchical Bloom levels to map interactions onto cognitive stages. Peng et al. (2019) proposed a conceptual framework for personalized adaptive learning environments utilizing Bloom's Taxonomy, arguing that tailoring instruction to learners' cognitive levels could enhance comprehension and engagement. This is a design proposal rather than an empirical outcome trial. This observation aligns with broader findings in adaptive learning research in computer science education (Barbosa et al., 2024). Mousavinasab et al. (2021) conducted a descriptive systematic review of intelligent tutoring systems in STEM education, reporting that personalized feedback loops and adaptive scaffolding were commonly featured design elements across the reviewed studies. The review itself does not establish a consistent causal effect on learning outcomes across those studies.

2.3. *Role of Visualization in Learning*

Research has demonstrated that visualizations and animations can significantly enhance students' understanding of algorithms and data structures (Hundhausen et al., 2002). Interactive visualizations make abstract processes tangible, helping learners track changes

in real time and reducing cognitive load. Hundhausen et al. (2002), in a meta-analysis of 24 experimental studies of algorithm visualization, found that visualization technologies were associated with improved learning outcomes primarily when learners actively engaged with the visualizations (e.g., through construction or prediction) rather than passively viewing them; effects on cognitive load and task efficiency varied across the reviewed studies rather than reflecting a single experiment. Similarly, Kogan et al. (2024) found that students who used a visualization tool scored higher on exams and reported better usability experiences compared to those who used traditional methods. However, visualization alone may create an illusion of understanding. Research shows that visualization is most effective when learners actively interact with it. In addition, studies indicate that embedded questions and input manipulation lead to stronger conceptual understanding than passive animation viewing (Naps et al., 2002).

2.4. *Gamification and Motivation*

Motivation is a key factor for successful learning, especially in self-directed or technology-based environments. However, many existing algorithm visualization tools focus primarily on delivering content and lack gamification elements that can sustain user engagement over time. Gamification refers to the integration of game-like elements, such as points, badges, levels, challenges, time-based tasks, and feedback loops, into non-game environments to promote motivation and behavioral persistence.

Studies have shown that gamification can improve cognitive engagement, emotional connection, and task completion rates in educational contexts. For example, Dicheva et al. (2015) reviewed gamification in education and found that incorporating gamified strategies in learning platforms was associated with increased learner motivation and involvement across the reviewed studies, and Oliveira et al. (2026) subsequently reported controlled experimental evidence of substantial performance gains from well-integrated gamification in higher education.

Toda et al. (2017) conducted a systematic mapping study on the negative effects of gamification in education. Their study argues that gamification design can produce negative impacts in educational contexts and should therefore be applied carefully. Bai et al. (2020) conducted a meta-analysis of empirical research on gamification in higher education and reported moderate improvements in student engagement, but comparatively smaller gains in learning outcomes. These results highlight the need to carefully integrate gamification elements with instructional objectives, rather than employing them as superficial motivational enhancements.

In a randomized-controlled trial, Coelho et al. (2026) found that a combination of points, badges, and challenges produced significantly greater learning gains relative to a no-gamification control.

In the context of algorithm learning, gamification can transform passive viewing of visualizations into active problem-solving. Features such as unlocking new levels after mastering a topic, earning rewards for correct answers, and receiving immediate feedback may enhance the sense of progression and achievement.

2.5. Related Algorithm-Learning and Adaptive Systems

Several systems were developed for algorithm visualization. TRAKLA2 (Malmi et al., 2004) and OpenDSA (Fouh et al., 2014) are widely used web-based systems that combine automatically assessed algorithm-simulation exercises with visualization content. Perhák and Šimoňák (2022) describe a web-based system with three visualization modes – pre-defined visualizations, interactive visualizations, and interactive exercises – for visualizing algorithms and data structures. Grivokostopoulou et al. (2017) describe adaptive e-learning environments for programming and algorithms that adjust content sequencing to learner performance. In comparison to the systems overviewed above, the system developed for this study incorporates a Bloom's Taxonomy-based personalized prerequisite graph and lightweight gamification within a single bundled interface.

3. Research Objectives

The study aims to examine whether the complete adaptive, personalized, and gamified VZOU package, evaluated as a bundled intervention rather than as separable components, is associated with post-test performance, perceived workload, and perceived usability, compared with a non-adaptive version and a control condition.

To address these goals, the study was guided by the following research questions:

- RQ1:** To what extent does the complete adaptive, personalized, and gamified VZOU learning path (evaluated as a single bundled intervention rather than as separable components), using a Bloom's Taxonomy-based performance measure, influence learning performance compared with a non-adaptive version and a control condition?
- RQ2:** Do users of the adaptive personalized version report a higher level of perceived usability?
- RQ3:** Is there a difference in perceived mental workload between students using the adaptive personalized learning path tool and those using the standard version?
- RQ4:** Is there a correlation between perceived usability, perceived mental workload, and post-test performance?

Based on these questions, this research tested the following hypotheses:

- H1:** Students using the complete adaptive, personalized, and gamified VZOU package, evaluated as a bundled intervention, will achieve higher post-test performance, across Bloom's Taxonomy, compared to students in the non-adaptive or control conditions.
- H2:** Students who use the personalized learning path version of the VZOU tool will report higher levels of usability, compared to students in the non-adaptive or control conditions.
- H3:** Perceived mental workload will be lower among students using the personalized learning path version of the VZOU tool, compared to students in the non-adaptive or control conditions.

H4: Higher usability and lower perceived mental workload will positively correlate with improved post-test performance.

By answering these questions, the study aims to advance understanding of adaptive learning technologies, support new approaches for teaching complex computer science topics, and inform decisions about the integration of personalized and gamified visualization tools in STEM education at scale.

4. Methods

In order to evaluate the effects of personalization and gamification on learning achievements, a dedicated tool (VZOU) was further elaborated. This section outlines the methodology employed to design, develop, and evaluate the experimental tool. It begins by presenting the pedagogical and cognitive principles that informed the design of the VZOU platform. It then details the study materials and instruments, including the platform architecture, performance assessments, and questionnaires. Finally, it describes the experimental design, study procedure, and data analysis methods.

4.1. Design Principles

The design of the personalized adaptive version of VZOU interface was intentionally designed according to principles derived from CLT and Bloom's taxonomy. Rather than presenting all content and functional options simultaneously, the system employs progressive disclosure based on an individualized learning graph. This design reduces extraneous cognitive load by limiting interface complexity at each stage, while aligning task exposure with graded cognitive levels (e.g., remembering, application, analysis). In doing so, the system regulates intrinsic load and supports germane processing, enabling learners to progress hierarchically through increasing levels of cognitive complexity without unnecessary overload.

First, the platform is structured around a personalized learning graph, which is generated based on an initial knowledge-assessment questionnaire of the Knowledge Mapping Test, where learners begin the initial assessment used to personalize their learning path. The test identifies the learner's Bloom's level for each topic and maps it onto the adaptive learning graph. Each node in the learning graph represents a specific algorithmic topic and is initially classified based on the learner's initial platform level, following Bloom's Taxonomy (Anderson and Krathwohl, 2001). The graph encodes the pedagogical dependencies between concepts, enforcing a structured, step-by-step learning sequence, as depicted in Figure 1. For example, *Topological Sort* relies on the traversal mechanics of *Depth-First Search (DFS)*, making an understanding of DFS a necessary foundation. This sequential structure ensures that learners build foundational knowledge before advancing to more sophisticated topics. The path is updated according to the learner's level of knowledge.

Second, the platform's design incorporates principles from Cognitive Load Theory. To reduce extraneous cognitive load, the interface is intentionally kept clean and focused,

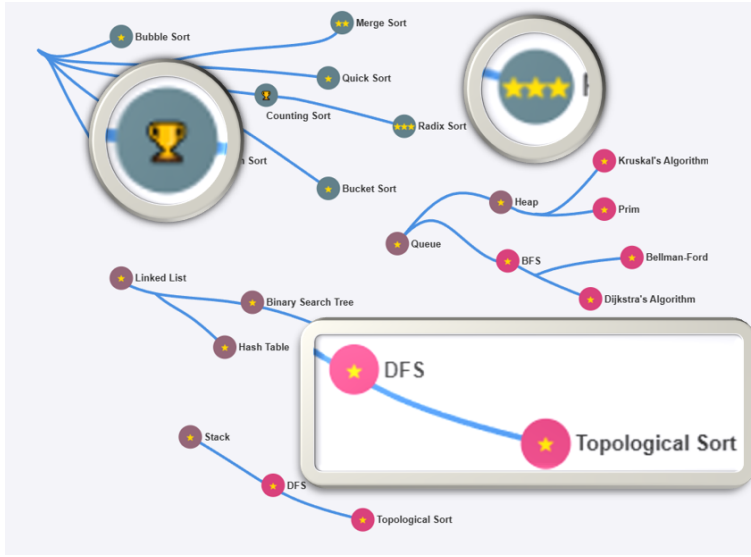


Figure 1. Zoomed-in view of the personalised learning graph—nodes represent algorithmic topics, and edges reflect prerequisite relations that guide structured learning.

with each visual module introducing one concept at a time. Animations and transitions are paced to support processing and retention without overwhelming the learner.

Third, the platform integrates basic gamification elements to maintain the engagement of the learners. For each completed challenge or concept node, learners receive visual rewards in the form of stars or trophies, that accumulate on their personalized progress graph. As shown in Figure 1, each node displays one to three stars or a trophy that reflects the learner's current progression state within the platform's Bloom-aligned learning sequence, providing immediate visual feedback and reinforcing motivation through a sense of achievement.

The adaptive sequencing logic governing node progression operates as follows. Each node in the learning graph is associated with three sequential quiz levels: ★ (Remember), ★★ (Apply), and ★★★ (Analyse). A learner's starting level for each node is set by the Knowledge Mapping Test (Section 4.2.1). Clicking a node initiates a multiple-choice quiz calibrated to the learner's current level for that topic. Successfully completing the quiz earns the corresponding star(s). If a quiz attempt is unsuccessful, the learner can always retake it. The same questions are presented on each attempt. A node displaying a trophy icon indicates that all three Bloom levels have been completed for that topic.

Crucially, unlocking a successor node requires earning at least one star (★) in each of its prerequisite nodes — that is, passing the Remember-level quiz is sufficient to open the path forward. Learners are not required to achieve ★★ or ★★★ before advancing through the graph, though they may return to complete higher levels at any time. This design balances structured progression with learner autonomy: it ensures foundational knowledge is confirmed before dependent concepts are introduced, while not artificially slowing learners who are ready to proceed.

These elements, along with real-time feedback and progress indicators, aim to maintain motivation and encourage continued engagement with the learning path.

4.2. Materials

This section details the core components used in the design and implementation of the study. It begins with a description of the tool development process, highlighting the technical aspects and features of the platform. The post-test subsection outlines the test questions, and the questionnaires subsection presents the instruments designed to collect user feedback and assess user experience. The section then moves on to the design and procedure of the experiment, which presents information about the participants, the experimental pipeline, and the experimental variables and conditions. Finally, the experiment procedure details the step-by-step process followed during the study.

4.2.1. Knowledge Mapping Test

In the adaptive learning version of the platform, the Knowledge Mapping Test is an assessment before any content is accessed. Its purpose is to establish each learner's current Bloom level for every algorithmic topic in the learning graph, which the system uses immediately to initialize the personalized learning path.

For each topic, the test presents two multiple-choice questions in sequence. The first targets the *Remember* level (★) and includes the opt-out option “I prefer to skip this topic,” allowing the learner to signal no prior familiarity with a topic entirely. The second targets the *Apply* level (★★) and includes the option “I prefer to skip this question,” allowing the learner to indicate conceptual familiarity but uncertainty about procedural execution. No *Analyze*-level (★★★) questions are included in the diagnostic; this level is reserved for in-platform progression. Table 1 presents representative items drawn from the test.

The initial Bloom level is assigned to each topic node based on the learner's responses using the following rule. If the learner skipped the topic or answered both questions incorrectly, the node is initialized at ★ (Remember). If the learner answered the Remember-level question correctly but skipped or incorrectly answered the Apply-level question, the node is initialized at ★★ (Apply), bypassing the Remember-level quiz requirement. If both questions were answered correctly, the node is initialized at ★★★ (Analyze), bypassing both lower levels. The resulting graph with its star indicators is visible to the user.

4.2.2. Tool Development

VZOU is an extension of the AVDS tool presented in Kogan et al. (2024). VZOU features an updated interface, extended algorithms visualization, and a newly built for the present adaptive version – the Knowledge Mapping Test, the personalized learning graph and gated progression logic, the formative-quiz system, and the gamification layer.

The VZOU platform was developed as a web-based system composed of three core layers: a client-facing React frontend, a Node.js-based RESTful API backend, and a PostgreSQL relational database. These components are deployed on a remote Windows server and communicate over secure HTTPS connections. Figure 2 illustrates the high-level system architecture.

Table 1
Sample Knowledge Mapping Test items by topic and Bloom level.

Topic	Level	Question	Options
Linked List	★ Remember	What is a Linked List?	A dynamic data structure with nodes linked together / A fixed-size array / A binary tree / I prefer to skip this topic
Bubble Sort	★ Remember	What does Bubble Sort do in each pass?	Finds the minimum element / Swaps adjacent elements if out of order / Divides the array in halves / I prefer to skip this topic
Bubble Sort	★★ Apply	After the first pass of Bubble Sort on [3, 2, 4, 1], what will the array look like?	[2, 3, 1, 4] / [3, 2, 4, 1] / [2, 3, 4, 1] / I prefer to skip this question
Insertion Sort	★ Remember	In Insertion Sort, where is the new element placed in each iteration?	At the beginning of the list / At its correct position in the sorted part / At the end of the list / I prefer to skip this topic
Insertion Sort	★★ Apply	Using Insertion Sort on [5, 2, 4], what will the array look like after inserting 2?	[2, 5, 4] / [5, 2, 4] / [2, 4, 5] / I prefer to skip this question
Stack	★ Remember	What is the main principle of a stack?	FIFO – First In First Out / LIFO – Last In First Out / Random Access / I prefer to skip this topic
Selection Sort	★★ Apply	Using Selection Sort, after one iteration on [4, 2, 1, 3], what will the array look like?	[1, 2, 4, 3] / [4, 2, 3, 1] / [2, 4, 1, 3] / I prefer to skip this question

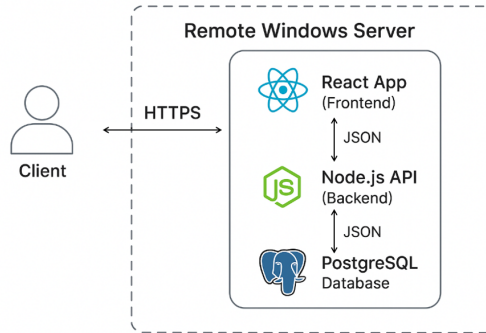


Figure 2. System architecture of the VZOU platform.

The system was designed with modularity, cross-browser compatibility, and future scalability in mind. Components were kept minimal and self-contained, following the KISS (Keep It Simple, Stupid) principle (Raymond, 2003) to ensure readability and ease of future extension.

The current version substantially extends that preliminary dissemination and prioritizes learner-centered adaptivity, shifting the system’s core focus toward dynamic instructional personalization. While the user-role and permission infrastructure remains embedded in the backend, administrative control is no longer the platform’s primary emphasis.

The VZOU platform evolved from an earlier system, referred to in this study as the “non-adaptive” version. This version operated without personalization capabilities and lacked key pedagogical and interface features, including the structured learning graph, the cognitively streamlined (low extraneous-load) interface, and gamification mechanisms that were subsequently incorporated into the adaptive framework. After gaining a deep understanding of the original architecture, the platform was significantly extended using React, a modern JavaScript library for building interactive user interfaces. While the original codebase was written primarily in TypeScript, the new components developed for the personalized learning graph and its animations were implemented in JavaScript, reflecting the developer’s greater fluency in that language.

The platform’s backend is built using Node.js with the Express framework, which handles server-side logic and data management. Communication between the frontend and backend is performed through RESTful API calls. These endpoints are used to submit the results of the initial knowledge-assessment questionnaire, as well as to update the learner’s current progress in the personalized graph. The backend processes this data to generate a tailored learning path and track ongoing performance.

User data, including topic-specific proficiency levels, is stored in a PostgreSQL relational database. Following the completion of the initial knowledge-assessment questionnaire, the backend computes the learner’s proficiency per topic and records it. This data serves as the foundation for constructing the personalized learning graph and is continuously updated as the learner progresses through the platform.

The platform follows a modular architecture, particularly in the visualization of algorithms. For each newly implemented algorithm, three dedicated components were developed: one for receiving user input, another for visualizing the algorithm’s behavior, and a third for managing communication between the two. This separation of concerns allows for better maintainability, scalability, and reusability of the codebase.

The personalized learning graph is rendered using a dynamic SVG object, which updates in real time as the learner progresses through the material. Node states and visual indicators are recalculated and redrawn each time a new concept is completed, providing an immediate and interactive representation of learning advancement.

Each node in the graph represents a specific topic and displays the learner’s current mastery using a star-based indicator system. As learners complete topics by successfully answering a topic-specific summary quiz tailored to their current Bloom level, they earn additional stars, reflecting advancement to higher levels of understanding. The graph enforces a strict sequential progression; nodes become accessible only after their prerequisite topics have been completed. This ensures structured learning and reinforces conceptual dependencies across the curriculum.

Each topic node in the learning graph serves as an interactive gateway to both assessment and visualization. Clicking on the node’s label redirects the learner to a dynamic

simulation of the corresponding algorithm, while clicking on the node itself initiates a knowledge-level-specific quiz. The visualizer component enables learners to input custom data and advance step-by-step through the algorithm execution. This interactive walk-through supports procedural understanding and bridges the gap between abstract concepts and concrete behavior.

The summary quizzes for each topic consist of multiple-choice questions designed to evaluate the learner's understanding at their current Bloom level. As the questions are closed-ended, the system does not provide full solution walkthroughs. Instead, feedback is limited to overall correctness, aligning with the quiz's function as a checkpoint rather than a teaching tool. To illustrate the quiz format across Bloom levels, Table 2 presents the in-platform quiz questions at the Analyze level (★★★) on Counting Sort topic. The Analyze level is the highest cognitive demand available in the platform. Questions at the ★ and ★★ levels follow the same multiple-choice format and are exemplified by the Knowledge Mapping Test items in Table 1.

Table 2
In-platform node quiz items at the Analyze level (★★★), topic: Counting Sort.

Question	Options
Why is Counting Sort not suitable for sorting the array $[-2, -1, 0, 1, 2]$ without modifications?	It cannot handle negative numbers without offsetting / It requires floating-point precision / It can only sort arrays of even length / It doesn't work for ranges below zero
How does Counting Sort achieve stability during sorting?	By placing elements in sorted order based on their cumulative counts / By rearranging elements randomly / By using additional memory to store temporary arrays / By performing in-place swaps
Compare Counting Sort to Quick Sort. Why is Counting Sort more efficient for sorting small-range data?	Counting Sort runs in $O(n)$ time for small-range data / Quick Sort is not stable for such data / Counting Sort uses divide and conquer strategies / Counting Sort requires no additional memory

Currently, the personalized learning graph serves as the main interface for visualizing learner progress. It reflects both the learner's progress through topics and their depth of understanding, using a graph layout and visual indicators. No additional summary tables, performance dashboards, or analytic views are provided at this stage.

The VZOU platform is freely available for students, educators, and researchers, and access details will be provided after acceptance. A demo of the platform is available at https://drive.google.com/file/d/1_jn4rePSsGD4sQ1vUK1XTf4ZZBD07ubX/view?usp=sharing.

4.2.3. Post-test

Participants completed a seven-item multiple-choice post-test designed to evaluate their understanding of two specific algorithms, Counting Sort and Radix Sort. All items were reviewed for alignment with the taught material and their assigned Bloom level by the course instructor (see Table 3). Four questions required simple fact retrieval (Remember), two questions asked learners to execute concrete algorithmic steps on short input lists (Apply), and one question checked the ability to analyze two steps of Radix Sort (Analyze). The responses were summed to score 0 to 100, based on the weighting scheme proposed in Table 4, in which questions associated

with higher Bloom taxonomy levels received greater weights. Higher scores reflect greater contributions from correctly answered higher-order questions.

The post-test was conducted digitally via Google Forms together with questionnaires. Responses were analyzed pseudonymously by mapping the email address to a random participant ID.

Table 3

Classification of post-test questions by Bloom's Taxonomy, with response options (correct option in bold).

Question	Bloom's Level	Options
Q1 – What is the primary limitation of Counting Sort?	Remember	It is not an in-place sort (it requires auxiliary arrays) / It is not efficient for arrays with a large number of unique values / It is not efficient for arrays whose value range is much larger than the array size
Q2 – What is the time complexity of Counting Sort?	Remember	$O(n^2)$ / $O(n \log n)$ / $O(n+k)$ / $O(\log n)$
Q3 – Which fundamental property of Counting Sort allows its use as a subroutine in Radix Sort?	Remember	Stability (it is a stable sort) / Minimal memory usage / Ability to handle negative numbers / Ability to sort real numbers
Q4 – In the time complexity expression $O(d \cdot (n + k))$ for Radix Sort, what does the variable d denote?	Remember	The additional memory required to run the algorithm / The maximum number of digits in the largest number in the array / The number of elements in the array / The number of auxiliary algorithms used in the sort
Q5 – In which case is Counting Sort inefficient?	Apply	When the value range is small relative to the array size / When the number of elements is very small / When the value range is significantly larger than the array size
Q6 – Following the summation step in Counting Sort, but before placement, what is the cumulative-sum array for the input array [4, 2, 2, 8, 3, 3, 1]?	Apply	[0, 1, 3, 5, 6, 6, 6, 7] / [1, 3, 5, 6, 6, 6, 6, 7] / [1, 2, 4, 5, 7, 7, 7, 7, 8] / [2, 4, 5, 6, 7, 7, 7, 8]
Q7 – For the array [55, 31, 97, 230, 451, 623, 142, 315], what is the intermediate order after sorting by the least and middle significant digits, prior to sorting by the hundreds digit?	Analyze	[315, 623, 230, 31, 142, 451, 55, 97] / [623, 315, 142, 230, 97, 55, 31, 451] / [31, 55, 97, 142, 230, 315, 451, 623] / [97, 55, 31, 230, 451, 315, 142, 623]

For each post-test question (Q1–Q7 in Table 3) we recorded a binary indicator (1 = correct, 0 = incorrect). To account for item difficulty (as per Bloom's taxonomy), we assigned each question a weight w_i :

The maximum possible sum of weights is $W_{\max} = 1 + 1 + 1 + 1 + 2 + 2 + 3 = 11$. Each participant's *WeightedRawScore* was computed as

$$\sum_{i=1}^7 (\text{Correct}_i \times w_i)$$

Table 4
Bloom-aligned item weighting used for post-test score aggregation. The weights reflect increasing cognitive complexity .

Question	Bloom's Level	Weight (w_i)
Q1, Q2, Q3, Q4	Remember	1
Q5, Q6	Apply	2
Q7	Analyze	3

and then normalized to a 0–100 scale:

$$\text{Score}_{0-100} = \frac{\text{WeightedRawScore}}{W_{\max}} \times 100.$$

The item weights were fixed prior to data analysis, based on each item's Bloom's taxonomy level as shown in the table above.

4.2.4. Questionnaires

To evaluate users' perceptions regarding the VZOU platform, we used two questionnaires: NASA-TLX (Hart and Staveland, 1988) to evaluate perceived mental workload and SUS (Brooke, 1996) for usability assessment.

The version of the NASA-TLX used in this study included six standard dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration. Participants rated each dimension on a scale from 1 ("very low") to 10 ("very high"), providing a quantitative assessment of the cognitive effort associated with using the platform.

Usability was assessed using the System Usability Scale (SUS) (Brooke, 1996), a widely adopted 10-item questionnaire designed to evaluate subjective system usability. Participants rated their agreement with each item on a 5-point Likert scale, ranging from "strongly agree" to "strongly disagree." The questionnaire included both positively and negatively worded statements, and scores were calculated according to the standard SUS method, resulting in a composite usability score between 0 and 100. Prior to analysis, an initial coding-direction error affecting all SUS responses was identified during data cleaning and corrected before computing scores using the standard SUS method (Brooke, 1996).

All questionnaires were administered digitally via Google Forms. Participants provided pseudonymization responses, and the collected data were used exclusively for research and evaluation purposes.

4.3. Participants

The study was conducted among eligible second-year students in the departments of Software Engineering and Computer Science at Sami Shamoon College of Engineering ($n = 55$). All participants were nearing the completion of a Data Structures course at the time of the study. Participants received a coupon for the café at the college as compensation for their participation. The course from which participants were drawn had 196 enrolled students. Each of the three groups was originally allocated 30 volunteer participants ($N = 90$ registered). However, 35 registrants did not attend the session or did not complete the assigned activity and instruments. The final analyzed sample ($N = 55$) consisted of 41 male and 14 female participants, aged 20–34 years ($M = 25.2$, $SD = 2.9$). All participants were second-year students who had completed the same prerequisite coursework and were enrolled in the same Data Structures course.

This study was approved by the Shamoon College of Engineering Human Research Ethics Committee (approval no. HREC-2025-0042, approved April 1, 2025). Recruitment and data collection took place over April–May 2025, just after the ethics approval. All participants gave informed, voluntary consent prior to participation; participation was independent of course grading, and participants could withdraw at any time without penalty.²

²An interim subset of this sample ($N=45$) was presented as an internal, non-peer-reviewed departmental showcase item at SCE TechFest 2025 (Data Structures and Algorithms Simulator), held July 30, 2025 (Bar, 2025).

4.4. Experiment Design and Procedure

This section outlines the experimental design used to evaluate the effectiveness of the VZOU platform in supporting algorithm learning. The study was designed to compare three learning conditions and assess their impact on students' comprehension, cognitive load, and perceived usability (i.e., traditional-control, non-adaptive and adaptive). To this end, participants were divided into three distinct groups, each experiencing a different version of the learning environment. The experimental pipeline is presented in Figure 3. We will overview each of the steps of the pipeline in Section 4.4.2.

4.4.1. Experiment Design

The study employed a between-subjects experimental design with three distinct groups of participants, each exposed to a different learning condition. Prior to registration, the research team randomly assigned each of the three parallel session slots to one of the three learning conditions (control, non-adaptive, adaptive VZOU). Participants then self-registered for a session slot without being informed which condition it represented. Assignment to condition was therefore randomized at the slot level and blind to participants at the point of registration. No individual-level randomization was performed, since the condition was fixed once a participant selected a (condition-blind) slot. Registration was unstratified, capped at 30 participants per slot ($N = 90$). Group 1 ($n = 16$) served as the control group and did not use any of the VZOU platforms, i.e., learned in the traditional way. Group 2 ($n = 17$) used the non-adaptive version of VZOU as a version of the digital platform without personalization features. Group 3 ($n = 22$) used the adaptive version of VZOU, which included an adaptive personalized learning graph based on Bloom's taxonomy, adaptive question flow, and gamified progress tracking.

4.4.2. Experimental pipeline

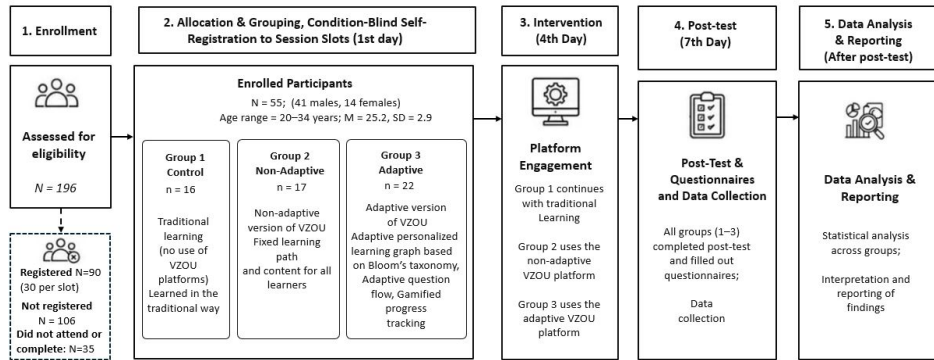


Figure 3. Overview of the experimental pipeline.

The experimental pipeline followed the stages presented in Figure 3.

Groups Setup. Participants self-registered equally to one of three condition-blind experimental groups (control, non-adaptive, adaptive). Each group received instructions according to its assigned learning condition.

Platform Engagement. Participants who used an application (i.e., Groups 2 and 3) were granted access on Day 4, three days after Day-1 registration and setup to their respective versions of the VZOU platform. They were instructed to explore the system independently and at their own pace to become familiar with its interface, features, and learning flow prior to the assessment phase.

Learning outcomes were assessed using a short-term post-test administered on Day 7 of the study.

Control (Group 1). Participants in the control condition (Group 1) did not access VZOU in any form. They were instructed to study the same topics: Counting Sort and Radix Sort, using the standard course materials available to all students enrolled in the Data Structures course: lecture slides, textbook chapters, and worked

examples provided by their instructors. They completed the post-test under the same time and conditions as the other two groups.

Non-adaptive VZOU workflow (Group 2). Participants in the non-adaptive condition accessed VZOU without the Knowledge Mapping Test, the personalized learning graph, or the star-based quiz system. All algorithm visualizers were accessible via free navigation learners could open any topic in any order without prerequisite constraints or adaptive sequencing. No formative quizzes were embedded in this version; the only formal assessment was the shared post-test administered on day 7.

Adaptive VZOU workflow (Group 3). Upon first login, participants completed the Knowledge Mapping Test (Section 4.2.1). On submission, the system generated the personalized learning graph, which was displayed at the platform's home screen. From this screen, learners could: (a) click a node label to open the interactive algorithm visualizer for that topic, enabling custom input and step-by-step execution; or (b) click the node icon to launch the level-specific quiz. Passing the quiz at a given Bloom level earned the corresponding star(s). Earning at least one star (★) in all prerequisite nodes unlocked successor nodes in the graph. Failed quizzes could be retaken immediately with the same questions. Platform access during the engagement period (Groups 2 and 3) was unsupervised and self-paced.

Post-Test and Questionnaires. Participants signed up for the Zoom session. All groups completed the NASA-TLX questionnaires immediately after the post-test, in the same procedural sequence. NASA-TLX ratings reflected the workload of completing the post-test after studying with standard course materials for the Control group, and the workload of using the respective VZOU version together with the post-test for the Non-adaptive and Adaptive groups. This learning activity constitutes the manipulated variable of this between-subjects design, as shown in the Intervention stage of Figure 3. The post-test and questionnaire procedure that followed (Figure 3, Day 7) was identical in sequence and timing across all three conditions. However, the SUS questionnaire was filled only by the adaptive and non-adaptive participants. All instruments were administered digitally, and each participant's test response time was recorded and analyzed as an additional dependent variable.

Data Collection. Data were collected using Google Forms and automatically recorded in associated Google Sheets. Each participant was required to enter their institutional email address to be correctly matched to their assigned experimental group. The email address was used solely as a participant identifier, and no additional personal information was collected. Each participant's institutional email address was linked to a randomly generated participant ID, that was used in the subsequent analyses. Six of the 55 submissions with incorrect or mismatched email addresses required correction (e.g., email typos, capitalization, or non-institutional email) to ensure data integrity and valid group attribution. Participants' privacy was preserved throughout this process, and was not linked to or accompanied by any further identifying or personal data.

Data collection was carried out over the course of one week during scheduled Zoom sessions. All responses were reviewed for completeness and consistency prior to analysis. Partial or incomplete submissions were excluded. Data were analyzed immediately after the final session, which formed the basis for evaluating the impact of the three different learning conditions on knowledge acquisition, cognitive load, usability, and task efficiency.

4.4.3. Data Analysis and Reporting

Following completion of the experiment, data were analyzed using SPSS version 31.00 to evaluate the effects of experimental condition on learning outcomes, cognitive load, usability, and task performance.

The independent variable was the assigned group condition (control, non-adaptive, or personalized adaptive). The dependent variables included:

- Post-test score (knowledge acquisition)
- Task completion time (in minutes)
- Cognitive load (NASA-TLX score)
- System usability (SUS score)

A one-way Analysis of Variance (ANOVA) was performed for the NASA-TLX cognitive load indices and post-test scores to assess differences between the three groups (Montgomery, 2020). The System Usability Scale (SUS) involved a comparison among the adaptive and non-adaptive environments, mean SUS scores were analysed with a one-way ANOVA, consistent with the analytic approach used for the NASA-TLX and performance measures. Homogeneity of variance was checked using Levene's test prior to comparison. Whenever the omnibus ANOVA yielded statistically significant results, post-hoc comparisons were carried out with Tukey-Kramer honestly significant difference (HSD) test (Abdi and Williams, 2010), which adjusts for the study's unequal group sizes, to identify the specific group differences. In addition to the overall test performance score, a supplement-

tary one-way ANOVA was conducted on post-test scores aggregated by Bloom's Taxonomy level (Remember, Apply, Analyze; Table 5), at the level of the individual participant.

The significance threshold for all tests was set at $\alpha = 0.05$. In addition, descriptive statistics, including means and standard deviations, were calculated to support the interpretation of the inferential findings. Post-test score is designated the primary outcome. Completion time, System Usability (SUS), the NASA-TLX composites and individual subscales, and the Bloom-level breakdown of post-test performance are secondary/exploratory outcomes. Tukey-Kramer HSD applies a family-wise correction to pairwise comparisons within each omnibus ANOVA. No correction is applied across the separate outcome families (primary outcome, secondary outcomes, correlational analyses), which we note as a limitation in Section 6.1. Given the primary outcome's small, unequal group sizes and its bounded 0–100 scale, Shapiro-Wilk tests and a Kruskal-Wallis rank-based sensitivity analysis were additionally conducted for the post-test-score ANOVA and for each Bloom-level ANOVA. Correlation and group-comparison analyses addressing NASA-TLX were conducted in two parallel forms. First, a six-sub-scale NASA-TLX total was computed as the mean of all six sub-scales (mental demand, physical demand, temporal demand, performance, effort, frustration), with the Performance sub-scale reverse-scored (reversed performance = $11 - \text{performance score}$) because its confirmed anchors, "I did not succeed" (1) to "I succeeded very much" (10), run in the opposite direction from the study's other five sub-scales. Second, because the Performance item's self-rated-success content is not fully independent of the objective performance outcome, we additionally computed a five-sub-scale composite that excludes the Performance item entirely (mental demand, physical demand, temporal demand, effort, frustration). Both composites were used (a) as the dependent variable in a one-way ANOVA comparing NASA-TLX workload across the three experimental conditions, and (b) in the RQ4 correlation analyses described below. In addition to these two composites, one-way ANOVAs were also conducted on each individual NASA-TLX subscale (mental demand, physical demand, temporal demand, performance/self-rated success prior to reverse-scoring, effort, and frustration) to identify which specific dimensions, if any, drove differences observed at the composite level. To address RQ4, we examined the association between post-test score (0–100), NASA-TLX workload (both composites), and SUS usability. Because SUS was administered only to the two VZOU arms (Group 2, non-adaptive, and Group 3, adaptive), and because experimental condition could confound any pooled association, we report: (1) the zero-order Pearson correlation between post-test score and each NASA-TLX composite across all three groups ($N = 55$); (2) a partial correlation controlling for condition; (3) descriptive within-group correlations for each condition (Control, $n = 16$; Non-adaptive, $n = 17$; Adaptive, $n = 22$); and (4) the correlations between SUS and post-test score and between SUS and NASA-TLX, restricted to the two VZOU arms ($N = 39$). All confidence intervals were computed using Fisher's z -transformation.

5. Results

The effects of personalization and gamification on the learning process are reported in the following section, divided into objective performance measures (post-test score and completion time) and subjective measures (System Usability Scale, SUS, and cognitive load, NASA-TLX) for the three experimental conditions: (1) control, (2) non-adaptive, and (3) adaptive.

Table 5 summarizes the descriptive statistics (mean $\pm$ SD) for the four key variables performance test score, completion time, SUS, and NASA-TLX across the three groups. SUS scores appear only for the adaptive and non-adaptive groups, because the control group did not interact with the tool; therefore, SUS is not applicable.

Post-test score. A one-way ANOVA found a significant effect of condition, $F(2, 52) = 6.93$, $p = .002$, $\eta^2 = .21$ (Levene's $F = 0.89$, $p = .415$; the Shapiro-Wilk test indicated non-normality, most pronounced in the adaptive group, $W = 0.76$, $p < .001$, so a Kruskal-Wallis sensitivity check was conducted and confirmed the result, $H(2) = 11.63$, $p = .003$). Tukey-Kramer HSD test showed that the adaptive group outperformed both the non-adaptive group ($p = .013$, 95% CI [2.94, 29.05]) and the control group ($p = .005$, 95% CI [4.95, 31.52]). The latter two groups did not differ ($p = .922$, 95% CI [−11.84, 16.32]).

The scores of the different groups varied across levels of Bloom's Taxonomy, such that the adaptive group scored highest at every Bloom level, with the largest numerical gap observed at the Analyze level as shown in Table 6.

Additional ANOVA analysis for performance was calculated, proportionally to the number of items at each Bloom's Taxonomy level (as shown in Figure 4). Scores multiplied by 100, converted every level onto

Table 5
Descriptive statistics for performance, usability, and cognitive load across experimental groups.

Variable	Adaptive (<i>n</i> = 22)	Non-Adaptive (<i>n</i> = 17)	Control (<i>n</i> = 16)
Post-test score (0–100)	89.26 ± 14.48	73.26 ± 19.41	71.02 ± 16.70
Completion time (min)	12.02 ± 4.78	11.69 ± 5.58	11.76 ± 6.04
System Usability (SUS)	75.80 ± 16.75	69.88 ± 20.11	N.A
Perceived Workload (NASA-TLX)	4.23 ± 1.24	4.04 ± 1.11	3.46 ± 0.98

Table 6
Performance scores (%) by Bloom's taxonomy level and condition.

Bloom Level	Adaptive (n=22)	Non-adaptive (n=17)	Control (n=16)
Remember	85.23 ± 23.98	80.88 ± 20.78	67.19 ± 26.95
Apply	88.64 ± 21.45	67.65 ± 30.32	71.88 ± 25.62
Analyze	95.45 ± 21.32	70.59 ± 46.97	75.00 ± 44.72

a common 0–100 percentage scale. Levene's test confirmed homogeneity of variance at all three levels (Remember: $F(2, 52) = 0.08$, $p = .926$; Apply: $F(2, 52) = 1.49$, $p = .234$; Analyze: $F(2, 52) = 2.42$, $p = .098$). Shapiro-Wilk tests indicated substantial non-normality at every level, consistent with the bounded, discrete nature of the underlying binary item scores. Therefore, Kruskal–Wallis sensitivity checks were conducted alongside each ANOVA. *Apply* was the only one of the three levels, that reached statistical significance, $F(2, 52) = 3.71$, $p = .031$, $\eta^2 = .12$. Tukey HSD post-hoc comparisons indicated that the adaptive group significantly outperformed the non-adaptive group at the apply level ($p = .037$). The control vs. adaptive and control vs. non-adaptive comparisons were not significant at this level. The *Remember* level questions, found approached but not significant $F(2, 52) = 2.731$, $p = .075$, and the *Analyze* level likewise did not reach significance ($p = .099$) despite showing the largest descriptive gap.

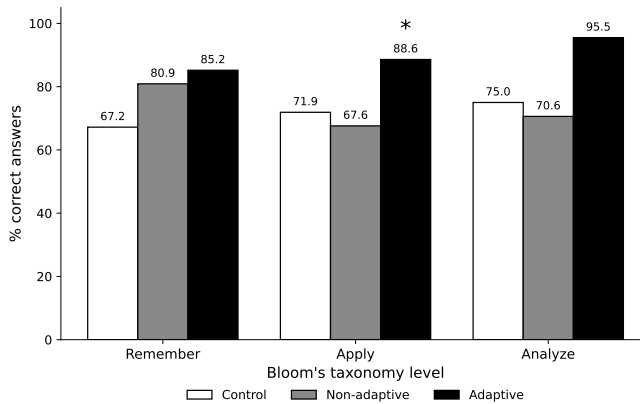


Figure 4. Percentage of correct answers by Bloom's taxonomy level for each group. * for $p < .05$

Completion time. A one-way ANOVA between-subjects did not show a reliable effect of the group on the duration of the task, $F(2, 52) = 0.021$, $p = .980$, $\eta^2 = .0008$. Levene's test confirmed homogeneity of

variance, $F(2, 52) = 0.235$, $p = .792$. Because Shapiro-Wilk indicated departures from normality in the Control and Non-adaptive groups (both $p < .05$), a Kruskal-Wallis test was also conducted and confirmed the null result, $H(2) = 0.350$, $p = .840$, confirming that task duration did not differ meaningfully across the three learning conditions.

System Usability (SUS). A one-way ANOVA comparing the non-adaptive and adaptive groups on the corrected SUS scores revealed no statistically significant difference, $F(1, 37) = 1.00$, $p = .323$, $\eta^2 = .03$. Levene's test confirmed the homogeneity of variance assumption was met, $F(1, 37) = 0.58$, $p = .452$.

NASA-TLX. The one-way ANOVA computed as the mean of the six subscales, with Performance reverse-scored per its confirmed anchors did not differ significantly across the three conditions, $F(2, 52) = 2.23$, $p = .118$, $\eta^2 = .079$. Levene's test indicated that the homogeneity-of-variance assumption was met, $F(2, 52) = 0.71$, $p = .496$. Tukey's HSD post hoc comparisons revealed no statistically significant pairwise differences: Adaptive versus Control, $p = .106$; Adaptive versus Non-adaptive, $p = .864$; and Control versus Non-adaptive, $p = .310$.

A parallel one-way ANOVA was computed on a five-subscale composite that excludes the Performance item entirely (mental demand, physical demand, temporal demand, effort, and frustration), given that the Performance item's self-rated-success content is not fully independent of the objective performance outcome (see Section 4.4.3). This composite differed significantly across the three conditions, $F(2, 52) = 3.33$, $p = .044$, $\eta^2 = .114$. Levene's test confirmed the homogeneity-of-variance assumption, $F(2, 52) = 0.21$, $p = .814$. Tukey's HSD post hoc comparisons showed that the adaptive group ($M = 4.58$, $SD = 1.47$) reported significantly higher workload than the control group ($M = 3.44$, $SD = 1.21$), $p = .036$. The adaptive-versus-non-adaptive comparison ($M = 4.25$, $SD = 1.34$ for the non-adaptive group) was not significant, $p = .729$, nor was the control-versus-non-adaptive comparison, $p = .213$.

At the individual-subscale level, Mental Demand differed significantly across conditions, $F(2, 52) = 5.78$, $p = .005$, $\eta^2 = .182$. Tukey's HSD post hoc test indicated a significant difference between the Adaptive and Control conditions ($p = .004$), whereas the difference between the Control and Non-adaptive conditions found marginally statistical significance ($p = .055$). Performance (self-rated success, prior to reverse scoring) showed a non-significant difference across conditions, $F(2, 52) = 2.94$, $p = .062$, $\eta^2 = .101$, with the Adaptive-versus-Control Tukey comparison at the conventional significance threshold ($p = .050$). Frustration also did not differ significantly across conditions, $F(2, 52) = 2.81$, $p = .069$, $\eta^2 = .098$. No individual Tukey comparison reached statistical significance, although mean Frustration scores increased monotonically from Control ($M = 1.62$, $SD = 1.09$) to Non-adaptive ($M = 2.82$, $SD = 1.67$) and Adaptive ($M = 3.23$, $SD = 2.81$).

No statistically significant between-condition differences were found for Physical Demand, $F(2, 52) = 0.59$, $p = .559$, $\eta^2 = .022$; Temporal Demand, $F(2, 52) = 0.78$, $p = .463$, $\eta^2 = .029$; or Effort, $F(2, 52) = 0.68$, $p = .513$, $\eta^2 = .025$. Levene's test confirmed homogeneity of variance for all subscales except Frustration, $F(2, 52) = 3.31$, $p = .044$. Therefore, the Frustration result should be interpreted with this caveat.

Correlation Analysis. Pearson correlation analyses assessed the relationships between performance, perceived usability (SUS), and perceived workload (NASA-TLX). Across all three groups ($N = 55$), the pooled correlation between performance and the six-sub-scale NASA-TLX composite was not significant, $r(53) = .241$, $p = .077$, 95% CI $[-.026, .476]$, controlling for condition, the partial correlation was $r = .164$, $p = .232$, 95% CI $[-.111, .416]$. The five-sub-scale composite excluding Performance showed a significant pooled correlation, $r(53) = .310$, $p = .021$, 95% CI $[.049, .532]$, which did not remain significant after controlling for condition, partial $r = .220$, $p = .106$, 95% CI $[-.053, .463]$. Within-group correlations were heterogeneous: non-significant in Control ($r = .156$, $p = .563$, 95% CI $[-.368, .605]$), significant and positive in Non-adaptive ($r = .682$, $p = .003$, 95% CI $[.300, .876]$), and negative and non-significant in Adaptive ($r = -.295$, $p = .182$, 95% CI $[-.637, .144]$). Within the two VZOU arms ($N = 39$), SUS was not significantly correlated with performance, $r(37) = .227$, $p = .164$, 95% CI $[-.095, .506]$ or with the six-subscale NASA-TLX composite, $r(37) = -.064$, $p = .699$, 95% CI $[-.372, .257]$.

6. Discussion

To frame the discussion, we revisit the research questions (RQs) together with the key results that address them.

RQ1. Students in the adaptive condition significantly outperformed those in both the non-adaptive and control groups on the post-test (ANOVA with Tukey). These results align with hypothesis H1. Because adaptivity and Bloom-sequencing are implemented as a single joint mechanism in this design rather than as separable features, this finding supports the complete adaptive package as evaluated, not adaptivity or Bloom-sequencing as isolable causal components. Such findings are consistent with recent controlled experimental evidence showing that well-integrated gamification can lead to substantial performance gains in higher education (Oliveira et al., 2026). The adaptive pathway likely facilitated concept comprehension and contributed to higher accuracy across Remember and Apply items. This design principle reflects a structured, rule-based sequencing approach in which an initial diagnostic sets each learner's starting point and subsequent progression is gated by quiz performance, consistent with scaffolding approaches that align task demands with learner readiness (Kalyuga and Sweller, 2005).

RQ2. SUS usability scores did not differ significantly between the adaptive and non-adaptive versions; H2 was therefore not supported. The numerically higher mean usability score in the adaptive group (75.8 vs. 69.88) is reported descriptively only and should not be read as evidence of improved usability.

RQ3. The overall NASA-TLX comparison was sensitive to how the Performance subscale was handled. The six-subscale total (Performance reverse-scored) did not differ significantly across conditions ($p = .118$), whereas a five-subscale composite excluding Performance entirely did ($p = .044$), with the adaptive group reporting higher workload than the control group. At the individual-subscale level, Mental Demand was significantly higher in the adaptive condition than in the control condition ($p = .004$). Taken together, H3 was not supported: the pre-specified primary NASA-TLX measure (the six-subscale composite, Performance reverse-scored) did not differ significantly across conditions. The adaptive condition showed significantly higher workload on the Mental Demand subscale and on the five-subscale composite excluding Performance. This pattern suggests, but does not establish, that the adaptive learning environment engaged deeper cognitive processes among learners, though not without additional perceived demand.

This elevation in self-reported mental demand, specifically, is compatible with, but does not confirm, a germane cognitive load interpretation Sweller (1988). Learners in the adaptive condition may have allocated more cognitive resources to schema construction and problem solving rather than being hampered by extraneous demands. However, this interpretation cannot be directly confirmed with the instruments employed in this study, as NASA-TLX measures overall perceived workload and does not decompose load into intrinsic, extraneous, and germane components.

Other explanations for the elevated mental demand should also be considered. The additional navigational steps in the adaptive interface may have increased perceived effort independently of deeper learning, representing extraneous rather than germane load. Alternatively, the system may have routed learners to more challenging content, increasing intrinsic load without necessarily promoting schema construction. Distinguishing among these explanations would require instruments designed to decompose cognitive load into its constituent types, such as the Cognitive Load Scale developed by Leppink et al. (2013), which separately measures intrinsic, extraneous, and germane load, unlike workload instruments such as NASA-TLX.

Previous research on effective learning indicates that deliberate cognitive effort and appropriately calibrated challenges can foster durable learning. Mental effort contributes positively when it represents meaningful cognitive involvement rather than disorientation or excessive cognitive load (Bjork et al., 2011). In line with this framework, our findings are compatible with the hypothesis that adaptive visualization environments channel cognitive resources toward more productive processing, although this cannot be established directly from NASA-TLX scores. Similarly, Paas et al. (2003) argue that productive effort, associated with a germane cognitive load, is a key indicator of meaningful learning.

The pattern observed in this study is compatible with previous literature, indicating that a higher demand in the adaptive condition is consistent with, but does not confirm, more constructive cognitive processing. A possible interpretation is that the adaptive system elicited more effortful cognitive engagement; however, alternative explanations cannot be ruled out on the basis of NASA-TLX scores alone, and the present data do not establish a causal link between elevated mental demand and the adaptive group's superior post-test scores. From a design perspective, these results underscore the value of tailoring difficulty in real time to maintain learner engagement.

RQ4. The pooled association between perceived mental workload and post-test performance did not reach statistical significance for either the full six-sub-scale composite or after controlling for experimental condition. Excluding the Performance sub-scale produced a similar pattern: although the pooled correlation was statistically

significant ($r = .310$, $p = .021$), it was no longer significant after controlling for condition. Within-group correlations were markedly heterogeneous and, critically, negative and non-significant within the adaptive group itself ($r = -.295$ for the full composite and $r = -.277$ for the composite excluding Performance). By contrast, the non-adaptive group showed a strong positive correlation ($r \approx .68-.73$). These findings do not support interpreting elevated mental demand as a within-group correlate of performance in the adaptive condition. SUS was not significantly associated with either performance or workload within the two VZOU arms (i.e group 2 and group 3) ($N = 39$). Therefore, H4 was not supported.

6.1. Limitations

Several limitations of this study should be acknowledged. First, our findings demonstrated that the adaptive intervention resulted in a significant performance advantage compared with the comparison conditions. While the apparent advantage at the analyze level, though descriptively large, should be interpreted with caution given that it is based on a single test item and did not reach statistical significance. Second, the relatively small sample size for each condition limits statistical power and may decrease the detectable magnitude of the adaptive system's effect compared to the non-adaptive and control conditions. Third, all participants were second-year students with the same prerequisites. Nonetheless, we recognize that individual variability in prior knowledge cannot be fully ruled out in the absence of a pre-test.

Fourth, because randomization was performed at the level of registration slots rather than individual participants, and participants self-selected into slots without knowing the associated condition, allocation is best described as slot-level and condition-blind rather than fully individually randomized. While this preserved blinding to condition at enrollment, it leaves open the possibility that unmeasured factors correlated with slot choice (e.g., participants' available time slots) could confound the observed group differences. Causal attribution to the adaptive package should be interpreted with this allocation structure in mind.

Fifth, the experimental setup occurred first. The intervention began later on the fourth day, and this setup-to-intervention interval was uniform across all three groups. The post-test was administered on Day 7, see Figure 3, which still constitutes a short-term measurement and does not inform about long-term retention. Future longitudinal studies will be required to examine sustained effects over time.

Sixth, cognitive load was assessed using NASA-TLX, a subjective workload instrument that does not decompose load into intrinsic, extraneous, and germane components Leppink et al. (2013). Future studies should employ instruments specifically designed for this purpose.

Seventh, no correction was applied across outcome families, only within each omnibus ANOVA via Tukey's HSD.

In addition, the adaptive condition was introduced as a bundled intervention. The personalized graph, gated progression, formative quizzes, gamification elements, and opportunities for repeated quiz attempts were introduced simultaneously. Consequently, the present design does not permit attribution of the observed performance differences to any individual component.

6.2. Recommendations for Future Work

Future work should focus on improving the interface of the adaptive platform to enhance usability without disrupting its core cognitive framework, which is built around an initial diagnostic assessment followed by a sequence of quizzes that adjust each learner's starting level and gate progression. Specifically, implementing guided navigation will help keep users on the optimal learning path, while real-time feedback mechanisms will reinforce correct responses and gently correct errors. At the same time, these design enhancements should be evaluated for their effect on germane cognitive load (Sweller, 1988) to support through personalized difficulty levels.

A promising direction for future development involves the integration of a personalized avatar system tightly linked to the existing cognitive framework. The learners would complete the diagnostic assessment and tiered quizzes to earn points, which they could use to unlock avatar customizations. By linking cosmetic rewards directly to pedagogical milestones, such as mastering specific Bloom taxonomy levels, this approach may improve motivation while, in principle, supporting germane cognitive load—that is, the mental effort that directly contributes to learning (Sweller, 1988). However, this study did not measure germane load directly, and this benefit remains a hypothesis for future testing.

Another important direction involves examining the long-term effects of adaptive learning environments on knowledge retention and familiarity with the system. Learning is a continuous process, and cognitive load may decrease as learners become more experienced and comfortable with adaptive features over time. Investigating these long-term dynamics could provide deeper insights into how adaptive interfaces support sustainable learning and engagement (Sweller, 2020).

7. Conclusion

The findings of this study show that the complete, bundled adaptive VZOU package, combining visualization, adaptive personalization, and gamification within a single integrated design, was associated with higher post-test performance in learning Data Structures and Algorithms, under the study's condition-blind, slot-level allocation. Perceived workload was higher on some measures, though not on the overall six-sub-scale composite. Usability showed no real gain. Because these elements are not implemented or evaluated as separable components, this is a property of the design as a whole, not of any individual element. Learners in the adaptive group reported higher mental demand, and higher overall workload once we excluded the self-rated success subscale. A pattern possibly consistent with germane cognitive load, though the six-sub-scale NASA-TLX composite itself was not significant once Performance was reverse-scored and included.

However, NASA-TLX does not directly decompose load types, and alternative explanations, including extraneous load from interface complexity or intrinsic load from increased task difficulty, cannot be ruled out. Because NASA-TLX provides subjective workload ratings and does not decompose load types directly, future work should incorporate instruments specifically designed to decompose cognitive load into its constituent types. For example, Cognitive Load Scale Leppink et al. (2013), which separately measures intrinsic, extraneous, and germane load, and which the authors of that scale explicitly distinguish from workload instruments such as NASA-TLX.

The findings suggest several practical implications for designing adaptive learning tools. First, presenting material in alignment with Bloom's taxonomy can help ensure that students master prerequisites before moving on to the concept that builds on them. Second, incorporating lightweight gamification elements, e.g., mastery-based rewards, may further reinforce progress. Third, to maximize their pedagogical value, educators should integrate the visualization tools in a structured manner and not solely rely on them as standalone assistance. Finally, designers should periodically monitor learners' perceived demand load to identify interface elements that unintentionally create friction, and refine the tool accordingly for a smoother learning experience.

Acknowledgments

The authors thank the students in the Data Structures course who participated in this study.

Data availability

The VZOU platform is publicly available at <https://ds-sce.sce-fpm.com/>

Ethics approval and consent to participate

The study was approved by the Shamoan College of Engineering Human Research Ethics Committee (protocol HREC-2025-0042, approved April 1, 2025). Participation was voluntary, and informed consent was obtained from all adult participants prior to the study; no minors were involved. Participants were informed they could withdraw at any time without consequence. Participants received a café coupon as compensation.

Generative AI disclosure

AI-assisted tools were used in preparing this manuscript. All outputs were checked by the authors, who remain responsible for the manuscript content: Grammarly (browser/document extension, used throughout manuscript preparation and revision, 2025–2026) for grammar verification, and ChatGPT (OpenAI, GPT-4/GPT-4o family, accessed intermittently during 2025–2026) for language editing suggestions.

Author contributions

Conceptualization: N.B., I.R., H.C. Methodology: N.B., I.R., H.C. Software: N.B. Validation: N.B., I.R., H.C. Formal analysis: N.B., I.R., H.C. Investigation: N.B., I.R., H.C. Writing – original draft: N.B., I.R., H.C. Writing – review and editing: N.B., I.R., H.C. Supervision: I.R., H.C. All authors read and approved the final manuscript.

References

- Abdi, H. and Williams, L. J. (2010). Tukey’s honestly significant difference (hsd) test. *Encyclopedia of Research Design*, pages 1–5.
- Anderson, L. W. and Krathwohl, D. R., editors (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom’s Taxonomy of Educational Objectives*. Longman.
- Bai, S., Hew, K. F., and Huang, B. (2020). Does gamification improve student learning outcome? evidence from a meta-analysis and synthesis of qualitative data in educational contexts. *Educational Research Review*, 30:100322.
- Bar, N. (2025). Data structures and algorithms simulator. Internal Book of Abstracts, SCE Tech Fest 2025, p. 159. Non-peer-reviewed departmental showcase abstract.
- Barbosa, P. L. S., Carmo, R. A. F. d., Gomes, J. P. P., and Viana, W. (2024). Adaptive learning in computer science education: A scoping review. *Education and Information Technologies*, 29:9139–9188.
- Bjork, E. L., Bjork, R. A., et al. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the real world: Essays illustrating fundamental contributions to society*, 2(59-68):56–64.
- Brooke, J. (1996). Sus: A quick and dirty usability scale. In Jordan, P. W., Thomas, B., Weerdmeester, B. A., and McClelland, I. L., editors, *Usability Evaluation in Industry*, pages 189–194. Taylor and Francis, London.
- Coelho, F., Rando, B., Aparício, D., Pontífice-Sousa, P., Gonçalves, D., and Abreu, A. M. (2026). The impact of educational gamification on cognition, emotions, and motivation: a randomized controlled trial. *Journal of Computers in Education*, 13:537–584.
- Dicheva, D., Dichev, C., Agre, G., and Angelova, G. (2015). Gamification in education: A systematic mapping study. *Journal of educational technology & society*, 18(3):75–88.
- Fouh, E., Karavirta, V., Breakiron, D. A., Hamouda, S., Hall, S., Naps, T. L., and Shaffer, C. A. (2014). Design and architecture of an interactive eTextbook — the OpenDSA system. *Science of Computer Programming*, 88:22–40.
- Grivokostopoulou, F., Perikos, I., and Hatzilygeroudis, I. (2017). An educational system for learning search algorithms and automatically assessing student performance. *International Journal of Artificial Intelligence in Education*, 27(1):207–240.
- Hart, S. G. and Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in Psychology*, volume 52, pages 139–183. North-Holland.
- Hundhausen, C. D., Douglas, S. A., and Stasko, J. T. (2002). A meta-study of algorithm visualization effectiveness. *Journal of Visual Languages & Computing*, 13(3):259–290.
- Kalyuga, S. and Sweller, J. (2005). Rapid dynamic assessment of expertise to improve the efficiency of adaptive e-learning. *Educational Technology Research and Development*, 53(3):83–93.
- Kogan, G., Chassidim, H., and Rabaev, I. (2024). The efficacy of animation and visualization in teaching data structures: A case study. *Educational Technology Research and Development*, 72:2349–2372.

- Leppink, J., Paas, F., Van der Vleuten, C. P. M., van Gog, T., and Van Merriënboer, J. J. G. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods*, 45(4):1058–1072.
- Malmi, L., Karavirta, V., Korhonen, A., Nikander, J., Seppälä, O., and Silvasti, P. (2004). Visual algorithm simulation exercise system with automatic assessment: Trakla2. *Informatics in Education*, 3(2):267–288.
- Mejeh, M. and Sarbach, L. (2025). Co-design: From understanding to prototyping an adaptive learning technology to enhance self-regulated learning. *Technology, Knowledge and Learning*, 30(2):1021–1046.
- Montgomery, D. C. (2020). *Design and Analysis of Experiments*. Wiley, Hoboken, NJ, 10 edition.
- Mousavinasab, E., Zarifsanaiy, N., Kalhori, S. R. N., Rakhshan, M., Keikha, L., and Saedi, M. G. (2021). Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. *Interactive Learning Environments*, 29(1):142–163.
- Naps, T. L., Rößling, G., Almstrum, V., Dann, W., Fleischer, R., Hundhausen, C., Korhonen, A., Malmi, L., McNally, M., Rodger, S., and Velázquez-Iturbide, J. A. (2002). Exploring the role of visualization and engagement in computer science education. *SIGCSE Bull.*, 35(2):131–152.
- Oliveira, W., Dantas Scaico, P., Hamari, J., Soltiyeva, A., Brambilla, A., and MilHosseini, S. (2026). The effects of gamification on students' academic performance. *Technology, Knowledge and Learning*.
- Ortiz-Rojas, M., Chiluiza, K., Valcke, M., and Bolanos-Mendoza, C. (2025). How gamification boosts learning in STEM higher education: a mixed methods study. *International Journal of STEM Education*, 12:1.
- Paas, F., Renkl, A., and Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, 38(1):1–4.
- Peng, H., Ma, S., and Spector, J. M. (2019). Personalized adaptive learning: An emerging pedagogical approach enabled by a smart learning environment. *Smart Learning Environments*, 6(9).
- Perháč, P. and Šimonák, S. (2022). Interactive system for algorithm and data structure visualization. *Computer Science Journal of Moldova*, 30(1(88)):28–48.
- Raymond, E. S. (2003). *The Art of UNIX Programming*. Addison–Wesley.
- Sajja, R., Sermet, Y., Cwierny, D., and Demir, I. (2025). Integrating AI and learning analytics for data-driven pedagogical decisions and personalized interventions in education. *Technology, Knowledge and Learning*, pages 1–31. Advance online publication.
- Shute, V. J. and Towle, B. P. (2003). Adaptive e-learning. *Educational Psychologist*, 38(2):105–114.
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285.
- Sweller, J. (2020). Cognitive load theory and educational technology. *Educational Technology Research and Development*, 68(1):1–16.
- Toda, A. M., Valle, P. H. D., and Isotani, S. (2017). The dark side of gamification: An overview of negative effects of gamification in education. In Cristea, A. I., Bittencourt, I. I., Lima, F., and Isotani, S., editors, *Researcher Links Workshop: Higher Education for All*, volume 715 of *Communications in Computer and Information Science*, pages 143–156. Springer International Publishing, Cham.
- Vateha, F. and Šimonák, S. (2025). Comparative visualization of algorithms and data structures. *Computing and Informatics*, 44:336–365.
- Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal*, 45(1):166–183.