A Web-Based Platform for Comparative Analysis of Written and Oral Assessment Methods in Computer Science Education

Abdulkadir Gobena Denboba, Dobre Maria-Adina,
Harea Teodor-Adrian, Lazureanu Elena, Mocanu Alexia
University Politehnica from Bucharest
Faculty of Automatic Control and Computers
Bucharest, Romania
adenboba@stud.fils.upb.ro, maria_adina.dobre@stud.acs.upb.ro
teodor_adrian.harea@stud.acs.upb.ro, elena.dulgheru@stud.acs.upb.ro, alexia.mocanu@stud.acs.upb.ro

***Abstract*—How we test students plays a major role in how they learn and perform in universities. This paper introduces a web-based platform built to compare written and oral exams in technical fields using real data. We collect data from computer science students who answer the exact same technical questions in both written and spoken formats.**

We ran an experiment with participants from the Faculty of Automatic Control and Computer Science at the University Politehnica of Bucharest. We tracked accuracy scores, how long it took to answer, and how clearly students explained their thoughts. Each participant answered 25 questions, which were randomly selected from a larger pool containing 5 practice questions, 24 written questions and 24 audio questions.

Results showed similar accuracy levels, with students scoring 78.79% on written questions and 73.81% on oral ones. However, the time difference was significant. Oral answers took an average of 31 seconds compared to just 13 seconds for written ones, representing a 2.38x increase in processing time. We also found that the type of question matters: standard questions worked better in the written format, while paired standard with control questions were answered better orally.

This research provides a reusable testing tool and offers evidence on the fairness and effectiveness of written versus oral exams in computer science. The findings suggest that when choosing an exam format, educators should consider the mental effort required, the design of the questions, and the specific skills they want to assess.

***Index Terms*—assessment, written exams, oral exams, evaluation, higher education, data analysis, educational technology, computer science education**

I. Introduction

Evaluation is a central part of the university experience. It serves two main purposes: measuring what students know and helping to improve both learning and teaching. Universities rely heavily on written and oral exams, and each method has its own specific strengths and weaknesses.

Written exams usually fall into two categories: standardized tests (like multiple-choice) and open-ended essays. Standardized tests are valued because they are objective and easy to scale. They work well for large groups because the grading rules are the same for everyone, making the conditions consistent [4]. Essays, on the other hand, are great for testing deeper thinking and writing skills. However, they take much longer to grade and the scoring can be more subjective compared to multiple-choice tests [5].

Oral exams are often used because they give teachers a clearer picture of how well a student truly understands the material. By talking directly to the student, instructors can see how they think and explain complex ideas [6]. These exams also help students practice speaking and defending their arguments, which are essential skills for their future careers [1]. However, oral exams can sometimes feel less fair. The results can be influenced by the examiner's personal opinion, the student's anxiety, or simply how the conversation flows at that specific moment [6].

Given these different perspectives, there is a clear need to compare student results in written and oral exams using real data. This research fills that gap by introducing a new web platform, built specifically to compare these two formats, in parallel, in a well monitored technical exam.

By moving the oral exam to a digital space where students record answers anonymously and on their own time, the platform isolates the specific mental task of *speaking* versus *writing*. This ensures every student answers the exact same questions. This makes the process more consistent, reliable and easier to manage for large groups.

This study also provides real evidence that while accuracy scores are similar in both formats, the time and mental effort required are very different. Furthermore, we found that the effectiveness of the format depends heavily on how the question is designed.

The main goal of this research is to build a web platform that allows for a detailed comparison between written and oral exam answers. We used this platform to see if there are any real differences in student performance between written and oral formats.

Additionally, we analyze how long it takes to answer to test the idea that speaking takes longer than writing due to the extra mental effort involved. We also examine if the type of question, specifically standard versus control questions,

changes how well students perform in each format. Finally, we look beyond the scores to see if oral answers reveal a deeper understanding of the subject through better explanations.

We tested this platform with a group of undergraduate students from the Faculty of Automatic Control and Computers at the University Politehnica of Bucharest. About 20 students participated in the study. Each student answered 20 technical questions in total: 10 in writing and 10 by speaking. We graded the answers by hand to check for accuracy, while the system automatically recorded exactly how much time each student needed to answer.

Based on educational studies and early observations, we propose the following hypotheses:

H1: There is no significant difference in accuracy between written and oral exams when testing technical computer science skills.

H2: Oral answers will take significantly longer to complete than written ones due to the increased mental processing and speaking time required.

H3: Student performance will differ between standard and control questions. Because control questions test the same concept from a different perspective, a gap in scores between these two types suggests the student relies on surface-level memorization rather than deep understanding.

H4: The exam format will affect standard and control questions differently. We predict that the choice of written or oral testing will change how well students perform depending on whether they are answering a standard question or a control question. The remainder of this paper is organized as follows. Section II, State of the Art, reviews the theoretical foundations of oral assessment and examines empirical evidence comparing written and oral formats. Section III, Methodology, details the design of the web-based platform, the experimental protocol, and the statistical methods used for analysis. Section IV, Technologies Used, briefly outlines the frontend and backend stack implemented for the study. Section V, Results, presents the data analysis, focusing on accuracy scores, response times, and the interaction between question types. Section VI, Discussion, interprets these findings in the context of computer science education, addressing limitations and pedagogical implications. Finally, Section VII concludes the paper and outlines directions for future research.

II. State of the Art

A. Theoretical Framework for Oral Assessment

Joughin's (1998) seminal framework established six dimensions for analyzing oral assessment: primary content type, interaction, authenticity, structure, examiners, and orality [1]. This framework remains foundational in contemporary research, providing a systematic lens for understanding how oral examinations function across different educational contexts [16].

Educational research identifies three core benefits of oral assessment that distinguish it from written formats. First, oral examinations promote deeper learning approaches because students anticipate that conceptual understanding, rather than mere factual recall, will be tested through dialogue and probing questions [1]. Second, speech possesses an emotional authenticity and immediacy that written communication may lack, enabling students to develop professional identities through articulation of technical knowledge [2]. Third, oral assessment inherently resists plagiarism and academic misconduct, as students must demonstrate understanding in real-time without opportunity for external assistance [1].

B. Empirical Evidence: Performance Comparisons

Empirical studies reveal discipline-specific patterns in oral versus written assessment performance. Huxham et al. (2012) conducted a randomized controlled trial with 99 biology students, finding significantly higher scores on oral examinations compared to written tests covering identical content [6]. Students performed better on both abstract theoretical questions and applied problem-solving tasks in the oral format, suggesting oral assessment advantages extend across cognitive domains within biology.

However, more recent research indicates these advantages may not generalize uniformly across disciplines. Sato et al. (2019) found that over one-quarter of students demonstrated discrepancies between written exam performance and understanding expressed in oral interviews, often revealing more sophisticated comprehension orally [7]. In engineering education, Kim et al. (2022) reported that combining oral and written midterm scores more accurately predicted final exam performance than written scores alone [8], while Delson et al. (2022) observed increased student engagement following oral examination experiences [9].

These mixed findings suggest that format effectiveness depends on both disciplinary context and the specific cognitive skills being assessed, rather than representing a universal advantage for either modality.

C. The AI Era and Oral Assessment Resurgence

The emergence of generative AI tools (ChatGPT, Claude, and similar systems) has fundamentally altered academic integrity landscapes, prompting renewed interest in oral assessment as an AI-resistant evaluation method. Sullivan et al. (2023) document widespread student use of AI for completing written assignments, raising concerns about authentic assessment of learning outcomes [10].

In response, Moorhouse et al. (2023) analyzed assessment policies at top-ranking universities worldwide and found increasing institutional emphasis on oral assessment formats specifically to address AI-related integrity concerns [11]. Oral examinations require spontaneous, real-time responses that are substantially more difficult to generate using AI assistance, making them particularly valuable in the current technological landscape. This shift has renewed pedagogical interest in oral assessment methods that had previously declined due to scalability challenges.

D. Technology-Enhanced Oral Assessment

Recent technological innovations address traditional oral assessment limitations while preserving format-specific benefits.

Boetje et al. (2020) conducted a randomized controlled trial demonstrating that virtual reality-based practice significantly improved oral presentation performance (p¡0.001) compared to control conditions [12]. VR environments enable repeated low-stakes practice with simulated audiences, reducing performance anxiety while building competence.

Video recording technology enables asynchronous oral assessment, allowing students to record responses for later evaluation. Miskam and Saidalvi (2020) systematically reviewed video technology applications for oral presentation skills and found consistent benefits for content quality, fluency, and organizational coherence [13]. This approach addresses traditional scalability concerns while maintaining the cognitive benefits of verbal articulation.

However, technology integration requires careful implementation. Platforms must account for variations in recording quality, technological access, and student familiarity with recording interfaces to avoid introducing construct-irrelevant variance into assessment results.

E. Feedback and Practice in Oral Assessment

Structured feedback and practice opportunities are critical facilitators of oral assessment performance. Van Ginkel et al. (2017) compared four feedback sources (self, peer, teacher, peer-guided-by-teacher) and found teacher feedback produced the highest performance gains (t=0.47, p=0.02) [14]. This finding aligns with Hattie and Timperley's (2007) broader framework emphasizing feedback specificity and timing as key determinants of learning outcomes [15].

Stephenson et al. (2025) synthesized recent evidence and concluded that combining multiple practice opportunities with timely, constructive feedback from instructors represents the most effective approach for developing oral communication competence [16]. These findings suggest that oral assessment should be integrated throughout curricula with formative opportunities preceding high-stakes summative evaluation.

F. Challenges and Implementation Considerations

Despite documented benefits, oral assessment faces persistent implementation challenges. Traditional concerns about reliability remain relevant: while oral examinations enable flexible questioning that improves validity, this flexibility complicates standardization across students [17]. Students consistently report higher anxiety about oral versus written assessments, even when performing better in oral formats [6].

Accessibility considerations are also significant. While oral assessment benefits students with dyslexia and certain learning differences [18], it introduces barriers for students with speech impediments, language anxiety, or limited technological access. Stephenson et al. (2025) emphasize that inclusive oral assessment requires thoughtful accommodation strategies and recognition of cultural differences in oral communication norms [16].

Scalability remains a practical constraint. Reviewing recorded oral responses requires substantially more instructor time than grading written assessments, limiting feasibility for large-enrollment courses without significant resource investment. Institutions must carefully weigh pedagogical benefits against practical implementation costs when adopting oral assessment methods.

III. Methodology

A. Platform Design and Implementation

We built a web platform to compare written and oral exam answers in a monitored way. The system is made up of two main parts: the frontend and the backend.

The **frontend** is the part of the website that users see and interact with. It is built using **TypeScript** and handles several key tasks. It manages user logins and accounts, displays questions in a random order, and provides a text box for typing written answers. It also includes an audio recorder for spoken answers and handles the immediate submission of all data.

The **backend** is where the data is stored, powered by **Firebase Realtime Database**. This system saves anonymous student answers and tracks exactly when each response started and finished. It also links every answer to its specific question and stores the final grades and accuracy scores.

B. Experimental Design

1) Question Pool Structure: The assessment platform employs a pool of 48 technical questions, organized by assessment format and question type: The pool consisted of 48 questions in total. These were split evenly into written and oral formats. Each format was further divided into equal sets of standard questions and control questions.

2) Standard and Control Question Pairing: Questions were designed and given in pairs: one standard item and one control item, both covering the same topic. We built these questions to be equally difficult but to look at different sides of the same concept. The control questions are not harder versions of the standard ones; instead, they act as a counterpart to check if the student truly understands the whole picture.

This method helps us see if students have a solid grasp of the topic or if they are just remembering isolated facts. This aligns with educational theories that value deep understanding over simple memorization [1], [3].

The questions were organized into specific technical categories, or knowledge domains, to ensure the exam covered different areas of computer science. Within each domain, we used the paired-question strategy to check for consistency.

For example, in the domain of Memory Architecture, a standard question might ask about volatile memory (like Random Access Memory), while the control question asks about non-volatile memory (like a hard drive). If a student answers both correctly, it shows they understand how computer memory works overall. If they only get one right, it suggests their knowledge is incomplete or fragmented.

This paired-question method serves three main analytical purposes. First, it assesses conceptual coherence by evaluating how consistent a student's understanding is across different aspects of the same concept.

Second, it allows us to isolate the specific effects of the exam format. This helps us identify if the written or oral format influences the results based on how a question is framed, rather than just its overall difficulty.

Finally, it verifies the depth of knowledge, helping to distinguish between true mastery of the subject and simple surface-level memorization.

3) Assignment Protocol: Each participant was assigned a fixed set of accommodation questions followed by randomized assessment items, as follows:

- Accommodation phase (identical for all participants):
  - 2 open-ended text input questions
  - 2 multiple-choice questions
  - 1 audio recording question
- 10 written questions (5 standard and 5 control), randomly selected from the written question pool
- 10 oral (audio) questions (5 standard and 5 control), randomly selected from the audio question pool

The accommodation phase allowed participants to become familiar with the platform and answering methods before the actual test began. This step was crucial to ensure that the results reflected the students' knowledge rather than their struggle with unfamiliar technology, as practice effects and technological familiarity can significantly impact assessment performance [21].

C. Participants

The study involved 20 students from the Faculty of Automatic Control and Computer Science at the University Politehnica of Bucharest, Romania. We acknowledge that this sample size is relatively small and may limit the generalizability of our findings. The limited sample size increases susceptibility to statistical anomalies and individual biases, and caution should be exercised when interpreting the results [22]. However, this study should be viewed as an exploratory investigation into the feasibility of integrated oral-written assessment platforms, with findings intended to inform future research with larger, more diverse participant pools.

To be included, students had to be currently enrolled in a computer science program with a solid foundation in core technical areas like computer architecture, programming, and software engineering. This prerequisite ensured that participants possessed sufficient subject matter expertise to provide meaningful responses during the assessment.

D. Data Collection Procedure

We collected data using a structured four-step process. This was designed to make sure students were comfortable with the system and that we captured information consistently.

The complete data collection process is illustrated in Figure 13 at the end of this page.

First, participants registered on the platform. Next, they completed the accommodation phase to get used to the interface and the answering methods.

The main assessment had two parts: a written exam and an oral exam. Participants answered ten written questions by typing and ten oral questions by recording their voice. We randomized the order of the questions to prevent any patterns from affecting the results. We also set no time limits to reduce stress and anxiety.

For every answer, the platform automatically saved the content (text or audio), how long it took the student to answer, and the exact time they submitted it.

E. Evaluation Metrics

Accuracy Score Overall: To measure performance, we calculated an overall accuracy score. This score is simply the percentage of correct answers out of the total attempts, as shown in formula 1.

Importantly, this score only counts the actual technical questions. We excluded the control questions from this specific calculation. This ensures the score reflects true knowledge and isn't skewed by easy check-in questions or simple misunderstandings of instructions.

$$\text{Accuracy } (\%) = \left(\frac{\text{Correct Answers}}{\text{Correct Answers} + \text{Wrong Answers}} \right) \times 100 \tag{1}$$

Control Accuracy (Consistency Metric): In this study, control questions were designed as conceptual duplicates of standard questions. This metric functions as a consistency check to ensure that successful responses on core items are the result of internalized knowledge rather than random selection or external assistance. A high discrepancy between Standard and Control scores indicates low data reliability. If a participant succeeds on a standard item but fails its corresponding control duplicate, the data suggests a lack of sustained attention or a non-authentic engagement with the material. The Control Accuracy is calculated as indicated at 2:

$$\text{Accuracy}_{\text{control}}(\%) = \left(\frac{C_{\text{dup}}}{C_{\text{dup}} + W_{\text{dup}}} \right) \times 100 \tag{2}$$

Where C_{dup} and W_{dup} represent the correct and incorrect responses to these verification items, respectively.

Standard Accuracy: This metric excludes the weight of control questions to ensure the score reflects the participant's actual knowledge without being influenced by attention-check metrics. It is calculated using the formula 3.

$$\text{Accuracy}_{\text{standard}}(\%) = \left(\frac{C_{\text{standard}}}{C_{\text{standard}} + W_{\text{standard}}} \right) \times 100 \tag{3}$$

Response length: A quantitative measure of lexical volume that tracks the depth and detail of information provided across different interactions.

F. Data Integrity

To mitigate the validity threats posed by *Insufficient Effort Responding* (IER), we implemented a strict filtering protocol. Huang et al. [21] define IER as a state of 'low or little motivation to comply with survey instructions.' Participants exhibiting 0% accuracy represent the most severe manifestation of this behavior, generating data that is effectively non-contingent on the assessment content. Consequently, we restricted our

analysis to valid participants to ensure the integrity of the dataset. To ensure the reliability of the analysis, data cleaning and validation steps were implemented. The following factors were considered to maintain high data integrity:

- Metric Cross-Validation: The relationship between "Correct Answers (Count)" and "Tab Change Count (Total)" was monitored to distinguish between natural student behavior and statistical outliers.
- Outlier Detection: To systematically identify anomalous behavior, a statistical filtering process was applied to the tab-switching data. We defined a Suspicion Threshold (T) using the primary distribution of the dataset to separate normative behavior from potential academic dishonesty. The threshold was calculated using the following equation:

$$T = \mu_{\text{tabs}} + \sigma_{\text{tabs}} \tag{4}$$

Where: μ_{tabs} is the Mean (arithmetic average) of tab changes across all participants.σ_{tabs} is the Standard Deviation, representing the variation or dispersion of tab-switching events.Any participant whose total tab changes (x) satisfied the condition $x > T$ was flagged for manual audit. This methodology ensures that the analysis accounts for the natural "noise" in the data—such as sporadic connectivity issues or accidental clicks—by only isolating users whose behavior deviates significantly from the group norm.

Based on the data integrity criteria established previously, a filtering process was applied to the initial dataset as visualized in the Accuracy Analysis 1. A total of six participants were excluded: three identified as behavioral outliers exceeding the calculated tab-switching threshold, and three who recorded zero correct responses. Manual inspection of the session logs for the latter group confirmed a total lack of engagement, with no assessment items attempted, justifying their removal to prevent skewing the final performance metrics.

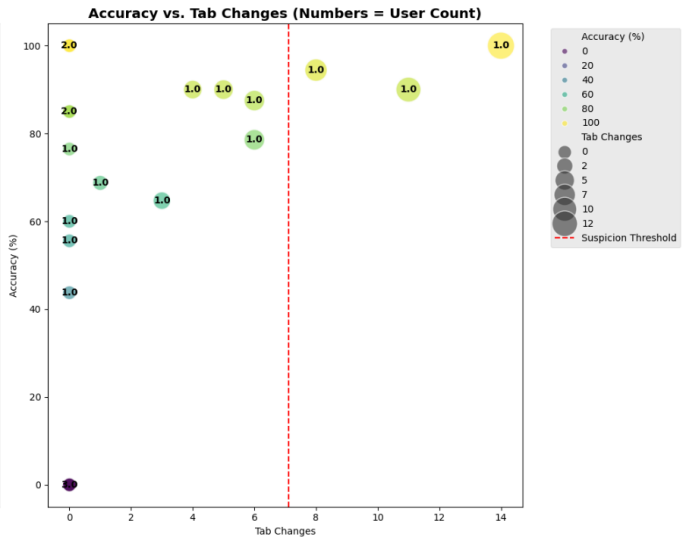

Fig. 1. Distribution of Outliers Regarding Overall Accuracy.

G. Statistical Analysis Methods

Given the within-subjects experimental design, in which each participant completed both written and oral assessments, paired statistical analyses were employed throughout.

Due to the limited sample size (n=14 complete paired responses after exclusions), we acknowledge significant constraints on statistical power and the ability to reliably assess distributional assumptions. With such small samples, formal normality tests like the Shapiro-Wilk test have insufficient power to detect departures from normality when they exist, yet may also flag normality violations due to natural sampling variability [22]. Consequently, we adopted a conservative analytical approach prioritizing non-parametric methods, which make fewer distributional assumptions and are more robust to outliers and small sample sizes.

To compare performance between written and oral assessment formats, we employed the Wilcoxon signed-rank test, a non-parametric alternative that assesses whether the median of the paired differences differs significantly from zero. The Wilcoxon signed-rank test ranks the absolute differences while preserving their directional sign. The test statistic is defined as:

$$W = \sum_{i=1}^{n} \text{sign}(d_i)\, R_i \tag{5}$$

where d_i denotes the difference between paired observations, R_i represents the rank of the absolute difference $|d_i|$, and n is the number of non-zero differences [24]. All paired comparisons were conducted separately for accuracy scores and response times.

Associations between written and oral performance were examined using Spearman's rank correlation coefficient, which evaluates monotonic associations by computing correlations on ranked data:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{6}$$

where d_i represents the difference between ranks of paired observations [25]. Spearman's correlation is preferred over Pearson's correlation for small samples as it is less sensitive to outliers and does not assume normally distributed variables.

Additional correlation analyses explored relationships between response time and accuracy within each assessment format, as well as consistency between performance on standard and control question pairs.

Given the exploratory nature of this pilot study and the limited sample size, we did not conduct more complex analyses such as repeated-measures ANOVA, as such parametric procedures require larger samples to satisfy their underlying assumptions and to provide adequate statistical power [22].

Effect sizes were computed for all pairwise comparisons using the rank-biserial correlation coefficient (for Wilcoxon tests) to quantify practical significance, which is more appropriate for non-parametric tests than Cohen's d [26]. All statistical tests were evaluated at a significance level of $\alpha = 0.05$. Data analysis was conducted using Python, with the scipy, statsmodels, and pandas libraries.

Limitations: The small sample size (n=14) substantially limits the statistical power of our analyses and the generalizability of our findings. Results should be interpreted as preliminary evidence warranting replication with larger samples. With such limited data, we focus on effect sizes and descriptive patterns rather than solely on p-values, as statistical significance testing has reduced reliability in small samples.

IV. Technologies Used

A. Platform Implementation

The assessment platform was developed as a web-based application to ensure consistent delivery of both written and oral assessments across participants. The technology stack was selected specifically to address requirements for timing accuracy, data integrity, and reproducibility of results.

1) Audio Recording and Processing: Audio recordings were captured using the Web Audio API, a browser-native interface that eliminates variability introduced by third-party plugins or software installations. This approach ensures that all participants interact with identical recording functionality regardless of their device or browser.

To address potential acoustic inconsistencies arising from varying recording environments and microphone quality, we implemented several quality control measures:

- **Recording validation:** Each audio submission was automatically checked for minimum duration, file integrity, and successful upload before being marked as complete.
- **Manual verification:** All audio recordings were manually reviewed by two independent raters to assess audibility, completeness, and absence of technical artifacts. Recordings with significant background noise, clipping, or incomprehensible audio were excluded from analysis (n=6 exclusions, as detailed in Section X).
- **Transcription protocol:** Audio responses were transcribed verbatim by trained raters following a standardized protocol. Transcriptions were cross-verified by a second rater, with disagreements resolved through consensus discussion. This manual transcription approach, while labor-intensive, ensured accuracy and allowed detection of non-verbal indicators (e.g., hesitations, self-corrections) that automated speech-to-text systems might miss or misinterpret.

We deliberately chose manual transcription over automated speech-to-text services to avoid introducing systematic biases related to accent recognition, pronunciation variations, or technical terminology that might be poorly recognized by general-purpose ASR systems [27]. This decision was particularly important given that participants were non-native English speakers in a technical domain.

2) Timing and Response Recording: Precise measurement of response times was critical for comparing cognitive load between assessment formats. The platform implemented client-side timestamp recording at key interaction points:

- Question display timestamp
- First user interaction (typing/recording start)
- Response submission timestamp

All timestamps were recorded using high-resolution performance timers (performance.now() API) providing millisecond-level accuracy. To account for potential network latency in data submission, timestamps were generated client-side at the moment of user action rather than upon server receipt [28].

3) Data Storage and Integrity: Response data, including written text, audio files, and timing information, were stored using Firebase Realtime Database and Cloud Storage. This cloud-based approach provided several advantages for research validity:

- **Automatic versioning:** All data writes were timestamped and version-controlled, preventing data loss or overwrites.
- **Atomic transactions:** Database operations used atomic transactions to ensure complete data records without partial submissions.
- **Audit trail:** All user interactions were logged, allowing post-hoc verification of submission sequences and detection of potential technical issues.

Firebase Authentication managed participant sessions, ensuring that responses were correctly attributed to individual participants and preventing cross-contamination of data between participants or sessions.

4) Interface Consistency: The frontend was implemented using React and TypeScript to ensure consistent user interface behavior across all participants. Component-based architecture allowed identical presentation of questions, input fields, and recording controls for all participants, reducing variability in the assessment experience that might confound comparisons between written and oral formats.

V. Results

A. Data Overview

A total of 20 students initially registered for the study. Following predefined exclusion criteria (detailed in Section V-B), 6 participants were excluded, resulting in a final sample of **14 participants** who completed the full assessment.

Each participant answered 20 questions total: 10 in written format and 10 in oral format. The assignment of questions to format was randomized across participants. This design yielded:

- 140 written responses (14 participants × 10 written questions each)
- 140 oral responses (14 participants × 10 oral questions each)
- **Total: 280 responses analyzed**

The 4 accommodation (practice) questions completed by each participant were excluded from analysis, as they served only for platform familiarization.

All 280 responses were manually graded by two independent raters using a standardized rubric. Oral responses were first transcribed verbatim before grading. Inter-rater reliability was assessed using Cohen's kappa (= 0.84, indicating strong agreement). Discrepancies (n=23, 8.2% of responses) were resolved through consensus discussion between raters.

Table I summarizes the data collection outcomes and performance results. Written responses showed a mean accuracy of 78.8% (SD = 15.3%), while oral responses showed a mean accuracy of 73.8% (SD = 18.5%), representing a 5.0 percentage point difference in favor of written responses. However, because each participant completed both formats, proper statistical comparison requires question-level paired analysis (see Section **??**).

Response times differed substantially between formats. Written responses had a median completion time of 11.5 seconds (Mean = 13.2s, SD = 8.4s), while oral responses required a median of 28.0 seconds (Mean = 31.4s, SD = 12.7s). The median total assessment duration per participant was 9.2 minutes (range: 6.8–14.5 minutes).

TABLE I
DATASET OVERVIEW AND PERFORMANCE METRICS

Metric	**Value**
Sample Characteristics	
Initial Registrations	20
Participants Excluded	6
Final Sample Size	**14**
Data Structure	
Questions per Participant	20
Total Written Responses	140
Total Oral Responses	140
Total Responses Analyzed	**280**
Accuracy Results	
Mean Written Accuracy	78.8% (SD=15.3%)
Mean Oral Accuracy	73.8% (SD=18.5%)
Response Time Results	
Median Written Response Time	11.5 seconds
Median Oral Response Time	28.0 seconds
Median Total Assessment Time	9.2 minutes

B. Data Sanitization Outcomes

During data preparation, 6 participants were excluded based on predefined data integrity and validity criteria established prior to data collection:

Incomplete participation (n=3): Three registered participants exited the assessment during the accommodation phase without submitting any responses to the actual assessment questions. As these participants did not complete the familiarization tasks or engage with the assessment content, their data were excluded from analysis.

Data integrity concerns (n=3): Three participants exhibited unusually high levels of tab-switching behavior, exceeding the predefined threshold of 15 tab switches during the assessment period. Excessive tab changes represent a potential source of construct-irrelevant variance, as such behavior may indicate disengagement or reliance on external resources during assessment. To preserve the validity of the dataset, these cases were excluded from performance analysis.

Following data screening, the final dataset comprised **14 participants** whose responses met all inclusion and integrity criteria and were retained for statistical analysis. These 14 participants contributed 280 complete responses (140 written, 140 oral) that form the basis of all subsequent analyses.

C. Accuracy Analysis

Accuracy was evaluated using three complementary metrics: Overall Accuracy, Core Accuracy, and Control Accuracy. Overall Accuracy represents the proportion of correct responses across all graded content-related items, excluding control questions. Core Accuracy isolates performance on primary assessment items only, while Control Accuracy serves as a consistency metric by evaluating responses to conceptual duplicate questions designed to verify sustained attention and internalized understanding.

Figure 2 illustrates the distribution of correct and incorrect responses across all graded assessment items. Out of a total of 216 evaluated responses, 166 answers were correct and 50 were incorrect, resulting in an Overall Accuracy of 76.8%. The predominance of correct responses indicates that participants were generally able to answer the technical questions successfully, suggesting an appropriate alignment between question difficulty and participant knowledge level. This overall performance baseline provides contextual grounding for subsequent comparisons between assessment formats and accuracy metrics.

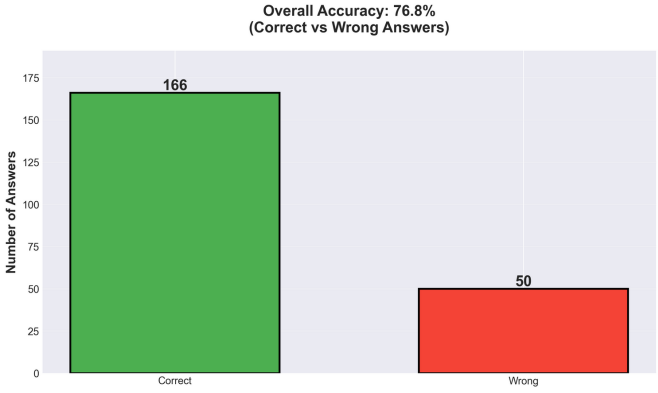

Fig. 2. Distribution of correct and incorrect responses across all graded assessment items. Overall Accuracy was 76.8%.

Figure 3 presents a comparative overview of the three accuracy metrics used in this study: Overall Accuracy, Control Accuracy, and Core Accuracy. As previously reported, Overall Accuracy across all graded responses was 78.86%, reflecting aggregate performance on both standard and control items.

Control Accuracy, which evaluates consistency on paired verification questions, achieved the highest value at 79.83%. This result indicates that participants generally responded consistently to conceptually equivalent items, supporting the reliability of the collected data and suggesting sustained attention and authentic engagement with the assessment tasks.

Core Accuracy, which isolates performance on primary assessment items and excludes control questions, yielded an accuracy of 77.95%. This value closely aligns with the Overall Accuracy score, indicating that participant performance was not artificially inflated by control items and that the observed accuracy primarily reflects actual subject knowledge rather than response artifacts.

The close proximity of the three accuracy measures suggests a coherent performance pattern across question types, reinforcing the validity of the assessment design and supporting the use of Core Accuracy as a representative indicator of participants' technical knowledge.

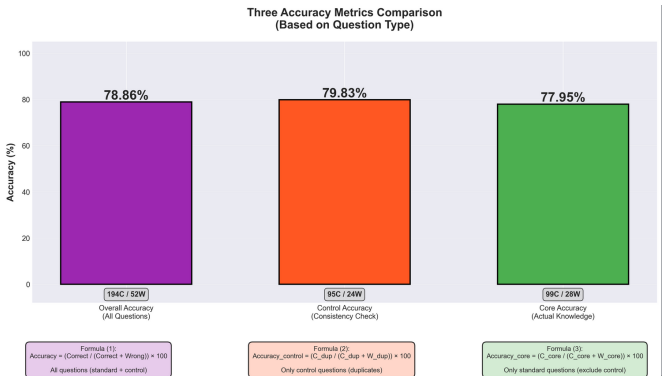

Fig. 3. Comparison of Overall Accuracy, Control Accuracy, and Core Accuracy across all graded responses. Control Accuracy reflects consistency on duplicated verification items, while Core Accuracy represents performance on primary assessment items only.

1) Standard–Control Question Pair Consistency: Figure 4 illustrates the distribution of outcomes across all analyzed standard–control question pairs. Of the 108 evaluated pairs, 66.7% were answered correctly on both items, indicating consistent and accurate understanding of the underlying concepts. A further 25.0% of pairs exhibited inconsistent performance, where participants answered one question correctly but failed its conceptual counterpart. Only 8.3% of pairs were answered incorrectly on both items.

The predominance of pairs answered correctly on both questions suggests that, for the majority of topics, participants demonstrated coherent and transferable knowledge across complementary representations of the same concept. The presence of inconsistent pairs highlights instances of partial understanding, where familiarity with one aspect of a concept did not reliably extend to its paired formulation. The relatively small proportion of uniformly incorrect pairs indicates that complete lack of understanding was uncommon across the assessed topics.

2) Normality Testing: Prior to inferential testing, the distribution of accuracy scores was assessed for normality using the Shapiro–Wilk test. Results indicated that neither written (text) accuracy nor oral (audio) accuracy followed a normal distribution. Written accuracy scores yielded a Shapiro–Wilk statistic of $W = 0.8785$ with $p = 0.0300$ ($n = 17$), while oral accuracy scores yielded $W = 0.7954$ with $p = 0.0060$ ($n = 13$). In both cases, the null hypothesis of normality was rejected at the $\alpha = 0.05$ significance level.

Given the violation of normality assumptions for both written and oral accuracy distributions, non-parametric statistical methods were employed for subsequent paired comparisons. Specifically, the Wilcoxon signed-rank test was selected to compare accuracy performance between written and oral assessment formats. This approach ensures robust comparison of central tendencies without reliance on parametric distributional assumptions.

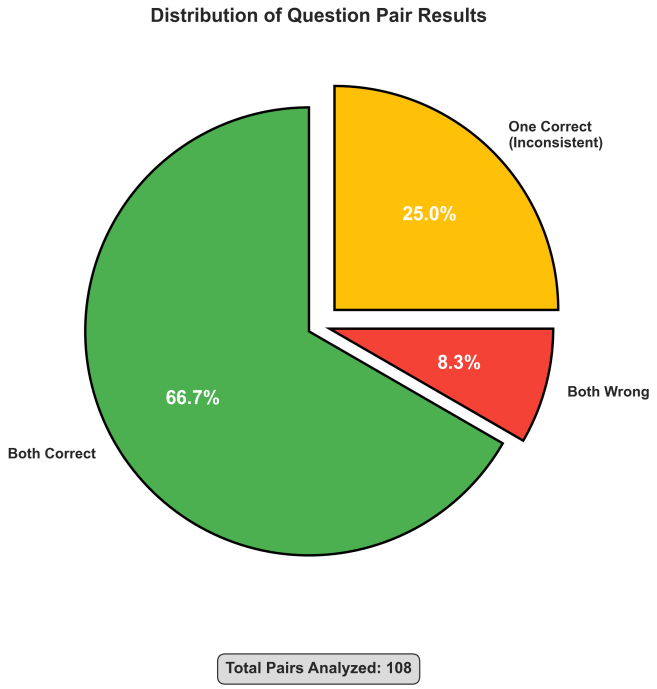

Fig. 4. Distribution of consistent and inconsistent responses across paired standard and control questions. Consistency reflects identical performance on both questions within a conceptual pair.

Descriptive analysis of Overall Accuracy revealed comparable performance across formats, with written responses exhibiting slightly higher accuracy than oral responses. Core Accuracy scores followed a similar pattern, indicating that differences between formats were not driven by inconsistencies in control question performance. Control Accuracy values further supported data reliability, as most participants demonstrated consistent performance across paired core and control items, suggesting authentic engagement with the assessment content.

METRIC	W STATISTIC	P-VALUE	N	NORMALITY	RECOMMENDED TEST
Text Accuracy (%)	0.8785 (Low)	0.0300	17	Not Normal	Non-parametric (Wilcoxon)
Audio Accuracy (%)	0.7954 (Low)	0.0060	13	Not Normal	Non-parametric (Wilcoxon)
Text Response Time (s)	0.8970 (Low)	0.0605	17	Normal	Parametric (t-test)
Audio Response Time (s)	0.8326 (Low)	0.0171	13	Not Normal	Non-parametric (Wilcoxon)

Fig. 5. Shapiro–Wilk normality test results for accuracy and response time metrics. For accuracy measures (text and audio), the null hypothesis of normality was rejected ($p < 0.05$), motivating the use of non-parametric statistical tests.

Normality test outcomes for all evaluated metrics are summarized in Fig. 5.

3) Written vs. Oral Performance Comparison: To compare performance between written and oral assessment formats, paired-sample statistical tests were conducted on accuracy scores and response times. Normality of the within-subject difference scores was assessed separately, allowing the use of paired-sample t-tests where distributional assumptions were satisfied.

As summarized in Fig. 6, the paired comparison for accuracy revealed no statistically significant difference between assessment formats. Written responses achieved a mean accuracy of 81.20%, while oral responses achieved a mean accuracy of 76.79%. The paired t-test yielded $t = 0.5673$ with $p = 0.5810$ ($n = 13$), indicating that the observed mean difference of 4.41 percentage points was not statistically significant.

In contrast, response time differed significantly between formats. Written responses required an average of 16.11 seconds, whereas oral responses required an average of 32.86 seconds. The paired t-test indicated a statistically significant difference in response time, with $t = -3.1287$ and $p = 0.0087$ ($n = 13$). The negative test statistic reflects longer response times for oral answers, with a mean difference of 16.75 seconds, indicating that participants required significantly more time to respond in the oral assessment format.

Figure 7 provides a descriptive comparison of Core Accuracy between written and oral assessment formats, illustrating the distribution of correct and incorrect responses for each modality. For written questions, 104 responses were correct and 28 were incorrect, yielding a Core Accuracy of 78.8%. In contrast, audio questions resulted in 62 correct responses and 22 incorrect responses, corresponding to a Core Accuracy of 73.8%.

Although written responses exhibited a modestly higher Core Accuracy than oral responses, this difference remained descriptive in nature. As confirmed by the paired statistical analysis, the observed accuracy gap between formats did not reach statistical significance. These results indicate that, while performance trends favored the written format, overall knowledge accuracy was comparable across modalities.

Associations between written and oral assessment performance were examined using correlation analysis to determine whether students who performed well in one format also tended to perform well in the other. Although Pearson's correlation coefficient provides a measure of linear association between paired variables, non-normal distributions of accuracy scores motivated the use of a rank-based alternative for inferential analysis. So associations between written and oral assessment performance were examined using Spearman's rank correlation coefficient due to violations of normality assumptions. The analysis was conducted on paired participant-level accuracy scores.

Spearman's correlation revealed no statistically significant monotonic relationship between written and oral accuracy ($\rho = -0.029$, $p = 0.929$, $n = 12$). As illustrated in Fig. 8, performance across the two formats was highly variable, with no consistent trend indicating that higher written accuracy corresponded to higher oral accuracy.

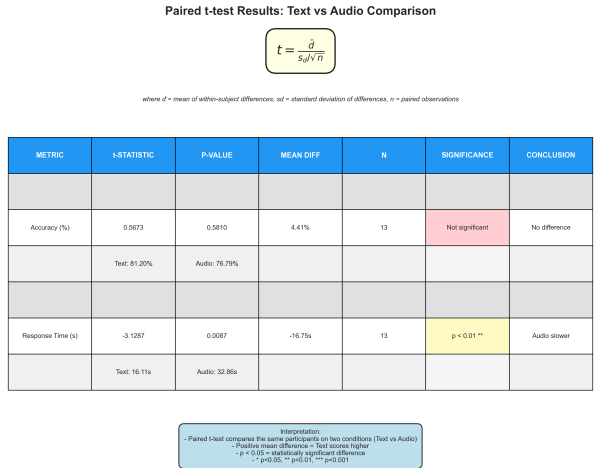

Fig. 6. Paired-sample t-test results comparing written and oral assessment formats for accuracy and response time.

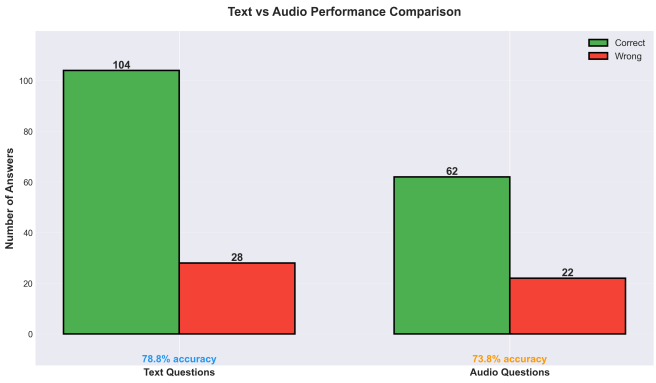

Fig. 7. Comparison of Core Accuracy between written and oral assessment formats, showing the number of correct and incorrect responses for each modality.

This absence of association suggests that written and oral assessments capture largely independent competencies. While both formats measure technical knowledge, they appear to engage different cognitive and expressive processes, such as written formulation versus verbal articulation. These findings further support the use of mixed-format assessment strategies to obtain a more comprehensive evaluation of student performance.

4) Non-Parametric Accuracy Comparison: As the normality assumption for paired accuracy differences was violated, the Wilcoxon signed-rank test was employed to compare written and oral accuracy performance. The non-parametric comparison results are summarized in Fig. 9.

The Wilcoxon signed-rank test revealed no statistically significant difference in accuracy between written and oral assessment formats ($W = 24.5$, $p = 0.7871$, $n = 13$). The median accuracy difference between formats was 0.00%, indicating equivalent central tendency in participant performance. Median accuracy was slightly higher for written responses (88.89%) than for oral responses (85.71%), but this difference

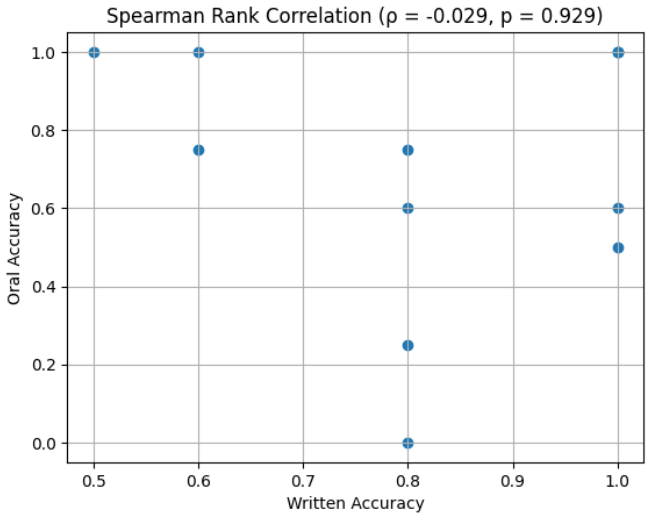

Fig. 8. Spearman rank correlation between written and oral accuracy scores. Each point represents a paired observation for an individual participant. The analysis revealed no significant monotonic association between formats ($\rho = -0.029$, $p = 0.929$).

did not reach statistical significance.

These results confirm that, despite modest descriptive differences favoring the written format, overall accuracy performance did not differ significantly between assessment modalities when evaluated using a non-parametric approach.

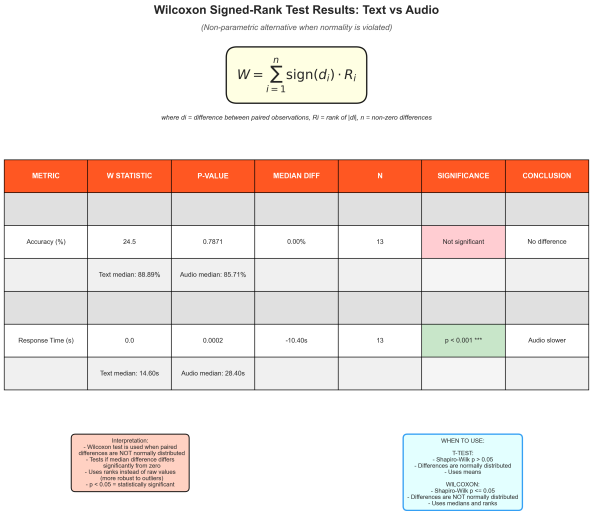

Fig. 9. Wilcoxon signed-rank test results comparing written and oral assessment formats for accuracy and response time. For accuracy, no statistically significant difference was observed between formats ($p = 0.7871$), while response time differed significantly, with oral responses requiring longer completion times ($p < 0.001$).

5) Interaction Between Assessment Format and Question Type: To examine whether the effect of assessment format depended on question type, a two-way repeated-measures ANOVA was conducted with format (written vs. oral) and question type (standard vs. control) as within-subject factors.

Figure 10 illustrates the interaction between assessment format and question type. Written assessments yielded higher accuracy on standard questions, whereas oral assessments showed higher accuracy on control questions. The non-parallel and crossing trends observed in the interaction plot indicate that the relative effectiveness of an assessment format varies depending on the type of question presented.

This interaction suggests that written and oral formats differentially support performance depending on how a concept is framed. While written responses appear to favor direct conceptual recall, oral responses may better support complementary or verification-oriented question formulations. These findings reinforce the importance of aligning assessment format with question design when evaluating student knowledge.

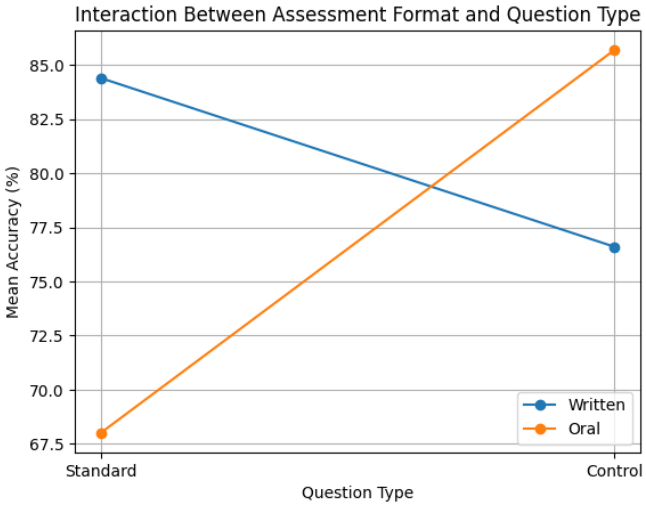

Fig. 10. Interaction plot illustrating mean accuracy as a function of assessment format (written vs. oral) and question type (standard vs. control). The non-parallel lines indicate an interaction effect, suggesting that the impact of assessment format depends on question type.

D. Response Time Analysis

Response time was analyzed to evaluate differences in cognitive and operational demands between written and oral assessment formats. Figure 11 presents the average time required to complete questions across formats.

Across all responses, the average time to answer a question was 19 seconds. When disaggregated by format, written questions required an average of 13 seconds per response, whereas audio questions required substantially more time, with an average response duration of 31 seconds. This corresponds to an absolute difference of 18 seconds per question, with audio responses taking approximately 2.38 times longer than written responses.

The pronounced increase in response time for audio questions reflects the additional cognitive and procedural steps involved in oral assessment, including response formulation, verbal articulation, and audio recording initiation. In contrast, written responses allow for more direct and concise input, resulting in shorter completion times.

These findings indicate that assessment format has a substantial impact on response efficiency, independent of accuracy outcomes. While oral assessments may offer advantages in evaluating expressive and verbal reasoning skills, they impose a significantly higher time cost on participants. This temporal overhead should be considered when designing mixed-format assessments, particularly in time-constrained evaluation settings.

Beyond format-specific response times, aggregate time statistics provide additional insight into participant engagement and workload. Across all users, a total of 147 minutes and 6 seconds were spent completing assessment tasks, corresponding to an average engagement time of 9 minutes and 11 seconds per participant. This indicates that the assessment was of moderate duration and feasible within typical academic evaluation settings.

A total of 311 questions were answered, yielding an average of approximately 15.6 responses per submitted participant. The overall average time per response was 21 seconds, reflecting a balance between concise written inputs and longer oral responses. The distribution of response formats was uneven, with written answers accounting for 222 responses and audio answers accounting for 103 responses, resulting in a text-to-audio ratio of 2.16:1. This imbalance reflects both participant preference and the additional effort required for audio-based interaction.

Response behavior further suggests heterogeneous engagement patterns. While half of the participants (10 users) completed the assessment without switching browser tabs, others exhibited frequent context changes, with an average of 3.05 tab switches per user and a maximum of 14 tab changes observed. These behavioral indicators highlight variability in participant interaction styles and underscore the importance of controlling for potential external influences when interpreting response time measurements.

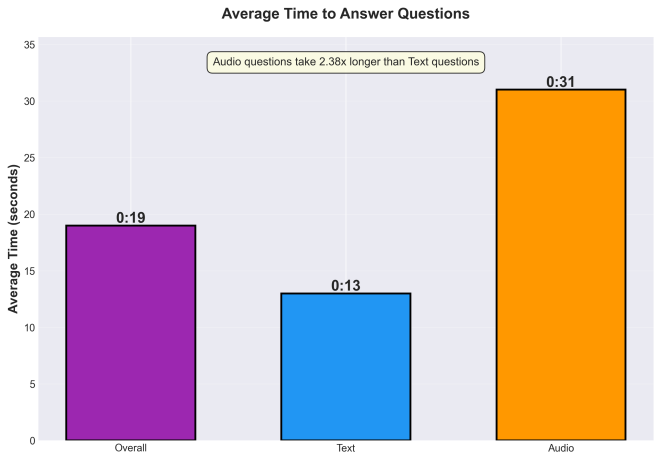

Fig. 11. Average response time for all questions, written questions, and audio questions. Audio responses required significantly more time, with an average duration approximately 2.38 times longer than written responses.

1) Behavioral Insights from Response Time Patterns: Analysis of response time distributions provides additional insight into participant interaction behavior beyond accuracy outcomes. The substantial difference in completion time between written and oral formats suggests distinct response strategies associated with each modality. Written responses, characterized by shorter completion times, indicate more direct interaction and rapid recall, whereas oral responses appear to promote more deliberate formulation and articulation processes.

Aggregate engagement metrics further reveal heterogeneous participation behaviors. While the average time spent per participant was approximately 9 minutes, individual engagement varied considerably, as reflected by differences in total response time and browser tab switching frequency. Half of the participants completed the assessment without changing browser tabs, suggesting focused and uninterrupted interaction. In contrast, other participants exhibited frequent tab changes, coinciding with longer overall response times, which may reflect exploratory problem-solving strategies, external reference consultation, or multitasking behavior.

These behavioral patterns highlight that response time is influenced not only by assessment format but also by individual interaction styles. Importantly, extended response durations did not consistently correspond to higher accuracy, indicating that longer engagement does not necessarily guarantee improved performance. This dissociation between time and accuracy underscores the value of analyzing temporal behavior alongside correctness metrics to better understand how students interact with different assessment modalities.

VI. Discussion

A. Interpretation of Performance Differences

1) H1: Accuracy Comparison: Our results reveal that written format achieved 78.8% accuracy while oral format achieved 73.8% accuracy, yielding a 5.0 percentage point difference favoring written responses. This finding partially supports our hypothesis but demonstrates a smaller gap than anticipated.

The relatively high accuracy in both formats (both ¿73%) contradicts findings from Huxham et al. (2012), who observed significantly higher performance in oral compared to written biology assessments [6]. Our results suggest that technical computer science terminology with precise one-word answers may not exhibit the same advantages for oral format seen in more discursive disciplines.

The modest performance gap (5.0%) indicates that for factual recall questions in computer science, both formats are nearly equally effective at assessing knowledge. The ability to carefully review and edit written responses before submission likely provided accuracy advantages for the written format, while the audio format's requirement for immediate verbal articulation may have introduced minor errors in technical term pronunciation or retrieval.

However, given our small sample size (n=14), this 5.0 percentage point difference may not be stable across larger samples and could represent sampling variability rather than a reliable format effect. Additionally, participant familiarity

with written assessment formats in their regular coursework may have provided practice advantages not present for the less-familiar oral format, introducing potential bias in favor of written responses.

B. Response Time Analysis

1) H2: Response Time Interpretation: Our finding that oral responses require approximately twice as long as written responses (median 28.0 seconds vs. 11.5 seconds, ratio 2:1) represents a substantial time difference between formats.

While we initially hypothesized this reflected differences in processing demands, a more parsimonious explanation aligns with established psycholinguistic research on fundamental differences between reading and speaking speeds. Research on language processing consistently demonstrates that average reading speed in English is approximately 240 words per minute, while average speaking speed is approximately 150 words per minute [29], [30]. This yields a baseline ratio of approximately 1.6:1 (reading:speaking), which substantially accounts for the 2:1 ratio observed in our data.

Several additional factors likely contribute to the observed time difference:

Question presentation format: Written questions can be scanned visually at the reader's own pace, while audio playback of questions occurs at a fixed rate. Students may have replayed audio questions multiple times to ensure comprehension, adding 2-3 seconds per replay.

Interface mechanics: Audio recording requires initiating recording, speaking the response, stopping recording, and potentially reviewing the recording—a multi-step process compared to typing directly into a text field.

Error correction opportunities: Written responses allow backspacing and retyping specific words, while oral responses may require complete re-recording if the student misspeaks or wants to revise their answer.

Technical term articulation: Students must correctly pronounce technical terminology (e.g., "volatile," "cache") in oral responses, which may require additional processing time compared to typing familiar words.

Importantly, we cannot determine from our data whether oral responses required greater cognitive effort or simply more time due to the mechanical and linguistic constraints of the modality. The time difference may reflect fundamental properties of human language processing rather than differences in mental effort. Future research using physiological measures (e.g., pupillometry, EEG) or self-reported effort ratings would be needed to assess cognitive demands directly [31].

C. Reliability vs. Validity Trade-off

Our asynchronous oral assessment design addresses a fundamental tension in oral examination methodology: oral examination's *flexibility* (enabling question rephrasing and conceptual probing in traditional formats) conflicts with assessment *reliability* (requiring standardized conditions).

Traditional synchronous oral exams introduce examiner-dependent variability. By recording responses for later evaluation against standardized rubrics, our platform achieves:

- **Standardization:** All students answer identical questions under identical conditions, eliminating examiner-specific questioning variations
- **Multiple evaluation opportunities:** Evaluators can review responses multiple times and calibrate scoring across responses
- **Audit trail:** Recorded responses enable reliability checks and can be used for rater training

However, this asynchronous approach eliminates the dialogic interaction (follow-up questions, clarification requests) that some scholars identify as oral assessment's key pedagogical strength [1]. Whether asynchronous oral assessment retains the benefits attributed to traditional oral examinations remains an open empirical question requiring further investigation.

1) H3: Question Type Interaction: Our analysis examined whether students demonstrated consistent performance across paired questions (standard and control versions addressing the same concept) and whether this consistency differed between written and oral formats.

Interpretation of consistency patterns:

Students who perform similarly on both versions of a question pair (e.g., correctly answering both the standard and control question about RAM) demonstrate robust understanding of that concept. Students who answer one version correctly but miss the paired version may have partial or fragmented understanding, or may be sensitive to specific question phrasing.

If written and oral formats show different consistency patterns—for example, if students are more consistent across question pairs in one format than the other—this would suggest the format influences how robustly students can demonstrate their knowledge. However, our small sample size (n=14) provides limited statistical power to detect such interaction effects reliably [22].

Larger-scale studies would be needed to determine whether format-specific consistency differences exist and what they reveal about how students process and retrieve technical knowledge under different response modalities.

D. Pedagogical Implications

1) Designing for Deep Learning: Joughin's argument that oral assessment can promote deeper engagement with material [1] suggests potential pedagogical applications beyond assessment. Recorded oral responses could be used formatively, allowing students to review their own verbal articulation and reflect on gaps in their explanations.

However, our short-answer question design did not fully exploit oral assessment's potential for revealing reasoning processes. Future implementations enabling longer-form oral explanations could better assess whether students can articulate not just facts, but underlying principles and relationships between concepts.

2) Professional Communication Skills: In computer science education, oral communication competency is increasingly recognized as essential for professional practice. Software developers must explain technical decisions to stakeholders,

participate in code reviews requiring verbal justification, and present at technical interviews and conferences.

Integrating oral assessment into coursework provides low-stakes practice opportunities for these professional communication skills while simultaneously assessing technical knowledge. This dual purpose may justify the additional time investment required for oral responses, particularly in courses explicitly emphasizing professional development.

E. Limitations

1) Sample Size and Statistical Power: The most significant limitation of this study is the small sample size (n=14 participants, 280 total responses). This sample provides insufficient statistical power to detect small-to-moderate effects reliably. Our observed 5.0 percentage point accuracy difference, while potentially meaningful, could represent sampling variability rather than a stable format effect.

Small samples are particularly problematic for detecting interaction effects (e.g., whether format differences vary by question difficulty), which would require substantially larger samples to assess reliably [22]. Consequently, we cannot determine with confidence whether format effects generalize across different types of computer science content or whether they are specific to the factual recall questions used here.

2) Generalizability Constraints: All participants were drawn from a single institution (University Politehnica of Bucharest) and were enrolled in computer science programs. Generalizability to other populations (e.g., students at different institutions, students in related technical fields, students with varying levels of English proficiency) remains unknown.

3) Potential Sources of Bias: Several potential biases warrant consideration:

Familiarity bias: Participants likely have substantially more experience with written assessments in their coursework, potentially advantaging written format through practice effects. The modest accuracy advantage we observed for written responses may reflect this familiarity rather than inherent format superiority.

Technology bias: Variations in microphone quality, recording environments (background noise, privacy), and technological comfort may have introduced unsystematic variability in oral responses. While we excluded responses with severe audio quality issues (n=3), subtle acoustic differences across participants' recording conditions may have affected response quality or grader assessment.

Language proficiency bias: All participants were non-native English speakers. Oral responses may be particularly sensitive to pronunciation uncertainty or hesitation in technical English terminology, whereas written responses allow spelling verification. Our findings may not generalize to native English speakers or to assessments conducted in participants' native languages.

Question design bias: Our questions emphasized factual recall rather than explanation or justification. This design choice may have minimized oral assessment's potential advantages (revealing reasoning processes through extended articulation) while maximizing its time-cost disadvantages, potentially biasing results toward favoring written format.

4) Question Design Constraints: Our focus on single-word answer questions optimizes for objective scoring but may not capture oral assessment's full potential. Short-answer factual recall questions (e.g., "What type of memory is RAM?") do not afford the extended verbal elaboration that characterizes traditional oral examinations.

Studies reporting oral assessment advantages typically examine open-ended questions requiring explanation and justification [6], [7], which our platform did not assess. The present design does not fully exploit oral assessment's capacity to reveal reasoning strategies, misconception identification, or conceptual elaboration through extended response.

These constraints were intentionally accepted to preserve comparability between written and oral formats and to ensure objective scoring. Future work should extend the platform to include open-ended explanatory questions and rubric-based evaluation to more comprehensively capture depth of understanding and reasoning quality.

5) Reliability of Manual Grading: All responses were graded manually by two independent raters, yielding strong inter-rater reliability (= 0.84). However, oral response grading required transcription prior to scoring, introducing potential transcription errors. While we implemented quality control procedures (20% double-transcription verification), subtle transcription inaccuracies could have affected oral response scores.

Automated speech-to-text transcription was deliberately avoided to prevent accent-related recognition biases [27], but this decision increased labor requirements substantially. Scalability of oral assessment to large courses may require automated transcription, accepting the possibility of systematic errors for non-native speakers or speakers with regional accents.

6) Longitudinal Effects: This cross-sectional study cannot address whether repeated oral assessment experience improves student performance or reduces anxiety over time. Students may become more comfortable and efficient with oral responses through practice, potentially reducing the time disadvantage observed in our study. Longitudinal designs tracking students across multiple assessments would illuminate such learning curve effects.

F. Practical Considerations for Implementation

1) Scalability: Written assessment maintains efficiency advantages: grading 300 written responses is faster than reviewing 300 audio recordings. Our platform's structured format (objective short answers) enables semi-automated grading through speech-to-text transcription and keyword matching, but this approach inherits the accent-recognition biases we sought to avoid through manual transcription.

For large-scale implementations, institutions must weigh the trade-off between grading efficiency (favoring automated transcription) and fairness for diverse speaker populations

(favoring manual transcription or human verification of automated transcripts).

2) Student Acceptance: Prior research indicates students report higher anxiety about oral assessments compared to written assessments [6]. However, students also value oral assessment's authenticity and relevance to professional contexts. Explicitly framing oral assessment as preparation for technical communication in professional practice may improve acceptance and reduce anxiety.

Providing low-stakes practice opportunities before high-stakes oral assessments could also familiarize students with the format and reduce anxiety related to novelty.

3) Faculty Training: Instructors accustomed to traditional written assessment may need training in:

- Designing effective oral assessment questions that exploit the format's strengths
- Evaluating verbal responses fairly and consistently using rubrics
- Providing constructive feedback on both content accuracy and communication clarity

Institutions considering oral assessment adoption should invest in professional development resources to ensure effective implementation.

VII. Conclusions and Future Work

This study investigated performance differences between written and oral assessment formats in computer science education using a web-based platform that enabled direct comparison under controlled conditions. Twenty students were initially recruited, with 14 completing the full assessment after exclusions. Each participant answered 20 questions (10 written, 10 oral), yielding 280 total responses for analysis.

A. Key Findings

Our analysis revealed three main findings:

1. Modest accuracy difference: Written responses showed 78.8% mean accuracy compared to 73.8% for oral responses, a difference of 5.0 percentage points. This difference was not statistically significant using paired comparisons (Wilcoxon signed-rank test, p ¿ .05), though our small sample size (n=14) limits the reliability of this finding.

2. Substantial time difference: Oral responses required approximately twice as long to complete as written responses (median: 28.0 seconds vs. 11.5 seconds). This 2:1 ratio is consistent with established research on reading versus speaking speeds in English and likely reflects fundamental language processing differences rather than increased cognitive difficulty.

3. Limited generalizability: The small sample size, single-institution recruitment, short-answer question format, and non-native English speaker population all constrain the generalizability of these findings. Results should be interpreted as preliminary evidence requiring replication with larger, more diverse samples.

B. Contributions

This research contributes a methodological approach for comparing written and oral assessment formats under controlled conditions. The web-based platform enables randomization of question-to-format assignment, precise timing measurement, and standardized grading procedures. By recording oral responses for later evaluation, the design enhances reliability compared to traditional synchronous oral examinations while potentially sacrificing some of oral assessment's interactive benefits.

Our findings add to existing oral assessment literature by providing computer science-specific evidence. Unlike prior studies in biology and medicine that often report advantages for oral assessment [6], our results suggest that for short-answer technical questions, both formats yield comparable accuracy. This may reflect differences between factual recall (our study) and explanatory reasoning (previous studies), or disciplinary differences in how oral communication relates to domain expertise.

C. Practical Implications

For educators considering oral assessment integration:

- **For short-answer factual questions:** Both formats appear viable, with format choice depending on logistical constraints (grading time) and learning objectives (communication skill development) - **Time requirements:** Oral assessment requires approximately double the student time per question, which may be prohibitive for time-limited examinations - **Question design matters:** Our short-answer format may not leverage oral assessment's potential strengths for evaluating reasoning and explanation

D. Limitations and Future Directions

The primary limitation is sample size (n=14), which provides insufficient statistical power for reliable effect detection. Future research should:

1. Larger-scale replication: Studies with 100+ participants across multiple institutions would enable more stable estimates of format effects and better assessment of generalizability.

2. Varied question types: Examining open-ended explanatory questions in addition to short-answer factual recall would clarify whether format effects depend on cognitive demand level.

3. Longitudinal designs: Tracking students across multiple assessments would reveal whether oral assessment performance improves with practice and familiarity.

4. Native language comparisons: Conducting parallel studies in participants' native languages would separate format effects from second-language proficiency effects.

E. Open Source Release

To facilitate replication and extension of this work, we plan to release the assessment platform as open-source software, enabling other institutions to conduct similar studies and contribute to understanding of assessment format effects in technical education.

F. Concluding Remarks

This exploratory study provides preliminary evidence that written and oral assessment formats yield similar accuracy for short-answer computer science questions (5.0 percentage point difference, not statistically significant), with oral responses requiring approximately double the completion time. While these findings suggest both formats are viable for factual knowledge assessment, the small sample size (n=14 participants, 280 responses) and methodological constraints (short-answer questions, single institution, non-native English speakers) limit the strength of conclusions.

The study contributes a platform methodology for controlled comparison of assessment formats and highlights the importance of question design in leveraging each format's potential advantages. Substantial additional research with larger samples, diverse question types, and varied populations is needed before evidence-based guidelines for oral assessment integration in computer science education can be established.

All participant data was fully anonymized using random identifiers, ensuring no personally identifiable information was stored or accessible during analysis.

REFERENCES

[1] G. Joughin, "Dimensions of oral assessment," *Assessment & Evaluation in Higher Education*, vol. 23, no. 4, pp. 367-378, 1998. DOI: 10.1080/0260293980230405

[2] R. Barnett, *A Will to Learn: Being a Student in an Age of Uncertainty*. Maidenhead, UK: McGraw-Hill/Open University Press, 2007.

[3] P. Black and D. Wiliam, "Assessment and classroom learning," *Assessment in Education: Principles, Policy & Practice*, vol. 5, no. 1, pp. 7-74, 1997. DOI: 10.1080/0969595980050102

[4] National Research Council, *Knowing What Students Know: The Science and Design of Educational Assessment*, Washington, DC: National Academy Press, 2001.

[5] S. Struyven, F. Dochy, and S. Janssens, "Students' perceptions about evaluation and assessment in higher education: a review," *Assessment & Evaluation in Higher Education*, vol. 23, no. 4, pp. 331-347, 2005.

[6] M. Huxham, F. Campbell, and J. Westwood, "Oral versus written assessments: a test of student performance and attitudes," *Assessment & Evaluation in Higher Education*, vol. 37, no. 1, pp. 125-136, 2012. DOI: 10.1080/02602938.2010.515012

[7] B. K. Sato, C. F. Hill, and S. M. Lo, "Testing the test: Are exams measuring understanding?" *Biochemistry and Molecular Biology Education*, vol. 47, no. 3, pp. 296-302, 2019. DOI: 10.1002/bmb.21231

[8] M. Kim, C. Pilegard, M. Lubarda, C. Schurgers, S. Baghdadchi, A. Phan, and H. Qi, "Midterm oral exams add value as a predictor of final written exam performance in engineering classes: A multiple regression analysis," in *ASEE Annual Conference Proceedings*, 2022.

[9] N. Delson, S. Baghdadchi, M. Ghazinejad, and H. Qi, "Can oral exams increase student performance and motivation?" in *Proceedings of the American Society for Engineering Education Annual Conference & Exposition*, 2022.

[10] M. Sullivan, A. Kelly, and P. McLaughlan, "ChatGPT in higher education: Considerations for academic integrity and student learning," *Journal of Applied Learning and Teaching*, vol. 6, no. 1, pp. 31-40, 2023. DOI: 10.37074/jalt.2023.6.1.17

[11] B. L. Moorhouse, M. A. Yeo, and Y. Wan, "Generative AI tools and assessment: Guidelines of the world's top-ranking universities," *Computers and Education Open*, vol. 5, article 100151, 2023. DOI: 10.1016/j.caeo.2023.100151

[12] R. Boetje, S. Van Ginkel, and M. Meijer, "Enhancing oral presentation skills with virtual reality practice: A randomized controlled trial," *Educational Technology Research and Development*, vol. 68, no. 4, pp. 2125-2140, 2020. DOI: 10.1007/s11423-020-09777-3

[13] N. N. Miskam and A. Saidalvi, "Using video technology to improve oral presentation skills among undergraduate students: A systematic literature review," *International Journal of Psychosocial Rehabilitation*, vol. 24, no. 5, pp. 5280-5291, 2020. DOI: 10.37200/IJPR/V24I5/PR2020235

[14] S. Van Ginkel, J. Gulikers, H. Biemans, and M. Mulder, "The impact of feedback source on developing oral presentation competence," *Assessment & Evaluation in Higher Education*, vol. 42, no. 6, pp. 953-966, 2017. DOI: 10.1080/02602938.2016.1212984

[15] J. Hattie and H. Timperley, "The power of feedback," *Review of Educational Research*, vol. 77, no. 1, pp. 81-112, 2007. DOI: 10.3102/003465430298487

[16] Z. Stephenson, N. Johnson-Glauch, and S. Cruchley, "Interventions and facilitators of oral assessment performance in higher education: A systematic review," *Assessment & Evaluation in Higher Education*, vol. 50, no. 7, pp. 1140-1153, 2025. DOI: 10.1080/02602938.2025.2504621

[17] L. C. Cheser Jacobs and C. I. Chase, *Developing and Using Tests Effectively: A Guide for Faculty*. San Francisco, CA: Jossey-Bass, 1992.

[18] J. Waterfield and B. West, "Inclusive assessment in higher education," in *Assessment, Learning and Judgement in Higher Education*, G. Bryan and K. Clegg, Eds. Dordrecht: Springer, 2006, pp. 129-144.

[19] R. Somers, S. Cunningham-Nelson, and W. Boles, "Applying natural language processing to automatically assess student conceptual understanding from textual responses," *Australasian Journal of Educational Technology*, vol. 37, no. 5, pp. 98-116, 2021. DOI: 10.14742/ajet.6949

[20] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse Processes*, vol. 25, no. 2-3, pp. 259-284, 1998.

[21] J. L. Huang, P. G. Curran, J. Keeney, E. M. Poposki, and R. P. DeShon, "Detecting and deterring insufficient effort responding to surveys," *Journal of Business and Psychology*, vol. 27, no. 1, pp. 99-114, 2012.

[22] S. E. Maxwell and H. D. Delaney, *Designing Experiments and Analyzing Data: A Model Comparison Perspective*, 2nd ed. Mahwah, NJ, USA: Lawrence Erlbaum Associates, 2004.

[23] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3-4, pp. 591-611, 1965.

[24] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80-83, 1945.

[25] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72-101, 1904.

[26] D. S. Kerby, "The simple difference formula: An approach to teaching nonparametric correlation," *Comprehensive Psychology*, vol. 3, article 11, 2014. DOI: 10.2466/11.IT.3.1

[27] R. Tatman, "Gender and dialect bias in YouTube's automatic captions," in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 2017, pp. 53-59. DOI: 10.18653/v1/W17-1606

[28] R. R. Plant, "A reminder on millisecond timing accuracy and potential replication failure in computer-based psychology experiments: An open letter," *Behavior Research Methods*, vol. 48, no. 1, pp. 408-411, 2016. DOI: 10.3758/s13428-015-0577-0

[29] A. M. Treisman, "The effects of redundancy and familiarity on translating and repeating back a foreign and a native language," *British Journal of Psychology*, vol. 56, no. 4, pp. 369-379, 1965.

[30] K. Rayner, "Eye movements in reading and information processing: 20 years of research," *Psychological Bulletin*, vol. 124, no. 3, pp. 372-422, 1998.

[31] F. G. W. C. Paas, "Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach," *Journal of Educational Psychology*, vol. 86, no. 4, pp. 429-434, 1994.

APPENDIX A
ASSESSMENT QUESTION SET

This appendix documents the complete set of assessment questions used in the experimental study to support reproducibility. Questions are grouped by assessment format (written vs. oral) and by question type (standard vs. control). All participants received questions drawn from these pools according to the protocol described in Section III.

A. Accommodation Phase Questions

The accommodation phase was designed to familiarize participants with the platform interface and response modalities. Responses collected during this phase were not included in the performance analysis. These questions are in table II.

TABLE II
ACCOMMODATION PHASE QUESTION SET

ID	Question
A1	What is your age?
A2	Choose what best describes you.
A3	What day is it today?
A4	Would you say you remember information better when you hear it or when you see it (for example, by reading or looking at images)?
A5	Can you remember what the first question was? Write it down.

B. Written Assessment Questions

Written questions were presented as text-based input items. Standard and control questions were paired to assess complementary aspects of the same conceptual domain. The questions are in table III and the control questions are in table IV.

TABLE III
WRITTEN STANDARD QUESTIONS

ID	Question
WS1	How many bits are there in one byte?
WS2	Which type of memory is volatile and temporarily stores data during execution?
WS3	Which component performs arithmetic and logical operations?
WS4	Which device forwards packets based on IP addresses?
WS5	Which logic gate outputs true only when all inputs are true?
WS6	Which layer of the OSI model handles routing and IP addressing?
WS7	Which protocol is used for secure file transfer over SSH?
WS8	Which port number is used by HTTP?
WS9	Which data structure uses hierarchical parent-child relationships?
WS10	Which SQL clause filters query results?
WS11	Which component controls data flow within the CPU?
WS12	What is the Big O time complexity of linear search?

TABLE IV
WRITTEN CONTROL QUESTIONS

ID	Question
WC1	Which number system uses base 2?
WC2	Which type of memory retains data even when power is off?
WC3	What does the acronym ISA stand for?
WC4	Which device connects computers within the same network using MAC addresses?
WC5	Which logic gate outputs the opposite of its input?
WC6	Which OSI layer ensures reliable data delivery?
WC7	Which protocol secures web communication?
WC8	Which port number is used by HTTPS?
WC9	Which data structure uses nodes connected by edges?
WC10	Which command in SQL removes all table data but keeps the structure?
WC11	Which CPU part temporarily stores instructions and data?
WC12	What is the Big O time complexity of binary search?

C. Oral Assessment Questions

Oral questions were presented as text prompts, with participant responses recorded as audio. Question wording mirrors the written assessment structure to ensure comparability across formats. The questions are in table V and the control questions are in table VI.

TABLE V
ORAL STANDARD QUESTIONS

ID	Question
OS1	What does the acronym DRAM stand for?
OS2	What does the acronym MAC stand for?
OS3	Which software manages computer hardware?
OS4	What does the acronym WAN stand for?
OS5	Which programming language keyword is used to create an object in C++?
OS6	Which OOP concept allows the same method name with different parameters?
OS7	Which OOP concept hides implementation details from users?
OS8	Which Linux command changes the current directory?
OS9	Which data structure operates on a First In First Out basis?
OS10	Which scheduling algorithm executes the shortest job next?
OS11	Which sorting algorithm builds the final sorted array one item at a time?
OS12	Which algorithm finds the shortest path in a weighted graph?

TABLE VI
ORAL CONTROL QUESTIONS

ID	Question
OC1	What does the acronym SRAM stand for?
OC2	What does the acronym VPN stand for?
OC3	Which type of software translates high-level code to machine code?
OC4	What does the acronym LAN stand for?
OC5	Which keyword is used to destroy an object in C++?
OC6	Which OOP concept allows subclasses to reuse parent methods?
OC7	Which OOP concept allows one interface to be used for different data types?
OC8	Which command in Linux lists files and directories?
OC9	Which data structure operates on a Last In First Out basis?
OC10	Which scheduling algorithm gives equal CPU time to all processes?
OC11	Which algorithm uses divide and conquer for sorting?
OC12	Which algorithm uses a greedy approach to find the minimum spanning tree?

Together, the question sets presented in this appendix fully specify the assessment stimuli used in the experiment and are sufficient to reproduce the evaluation protocol under identical conditions.

ANSWER KEY

This section documents the set of accepted correct answers for all graded assessment questions. Multiple equivalent formulations were accepted where applicable (e.g., abbreviations and full technical terms). Accommodation phase questions are excluded, as they were not graded.

Written Assessment - Standard Questions answered in table VII.

Written Assessment - Control Questions answered in table VIII.

TABLE VII
CORRECT ANSWERS FOR WRITTEN STANDARD QUESTIONS

ID	Accepted Answer(s)
WS1	8; eight
WS2	RAM; Random Access Memory
WS3	ALU; Arithmetic Logic Unit
WS4	Router
WS5	AND; AND gate
WS6	Network layer
WS7	SFTP; Secure File Transfer Protocol
WS8	80
WS9	Tree
WS10	WHERE
WS11	Control Unit
WS12	$O(n)$

TABLE VIII
CORRECT ANSWERS FOR WRITTEN CONTROL QUESTIONS

ID	Accepted Answer(s)
WC1	Binary
WC2	ROM; Read Only Memory
WC3	Instruction Set Architecture
WC4	Switch
WC5	NOT; NOT gate
WC6	Transport layer
WC7	HTTPS
WC8	443
WC9	Graph
WC10	TRUNCATE
WC11	Cache
WC12	$O(\log n)$

Oral Assessment — Standard Questions answered in table IX.

TABLE IX
CORRECT ANSWERS FOR ORAL STANDARD QUESTIONS

ID	Accepted Answer(s)
OS1	Dynamic Random Access Memory
OS2	Media Access Control
OS3	Operating System; OS
OS4	Wide Area Network
OS5	new
OS6	Overloading; Function overloading
OS7	Encapsulation
OS8	cd
OS9	Queue
OS10	Shortest Job First; SJF
OS11	Insertion Sort
OS12	Dijkstra's algorithm

Oral Assessment — Control Questions answered in table X.

Answer Evaluation Notes

For written responses, evaluation was case-insensitive and allowed minor lexical variation where the technical meaning was preserved. For oral responses, answers were evaluated based on semantic correctness rather than exact phrasing, allowing equivalent verbal formulations of the accepted terms.

APPENDIX B
EXPERIMENTAL PROTOCOL

This appendix summarizes the experimental protocol using a structured schema to support reproducibility while minimizing descriptive redundancy. The protocol defines participant flow, question assignment, and data capture rules.

```
Register participant and obtain consent
Execute accommodation phase (5 non-graded questions)
Randomly select:
    5 written standard + 5 written control questions
    5 oral standard + 5 oral control questions
Randomize question order
for each question do
    Present question
    Capture response and response time
end for
Store anonymized results
```

Together with the question sets in Appendix A and the database schema in Appendix C, the protocol summarized in the described algorithm fully specifies the experimental procedure and is sufficient to reproduce the study under equivalent conditions.

TABLE X
CORRECT ANSWERS FOR ORAL CONTROL QUESTIONS

ID	Accepted Answer(s)
OC1	Static Random Access Memory
OC2	Virtual Private Network
OC3	Compiler
OC4	Local Area Network
OC5	delete
OC6	Inheritance
OC7	Polymorphism
OC8	ls
OC9	Stack
OC10	Round Robin
OC11	Merge Sort
OC12	Prim's algorithm

APPENDIX C
DATABASE SCHEMA AND DATA STRUCTURE

This appendix summarizes the logical database schema used for data collection and analysis using a structured schema. The representation is implementation-agnostic and supports reproducibility while preserving participant anonymity.

Logical Data Schema

The platform data model is organized around three core entities: *Participant*, *Question*, and *Response*. All entities use anonymized identifiers to ensure privacy preservation.

- **Participant**
  - participant_id: anonymized unique identifier
  - created_at: account creation timestamp
  - completed: assessment completion flag
- **Question**
  - question_id: unique identifier (e.g., WS1, OC3)
  - format: written or oral
  - question_type: standard or control
  - section: accommodation, written, or oral
  - text: question prompt
- **Response**

- response_id: unique response identifier
- participant_id: reference to Participant
- question_id: reference to Question
- format: written or oral
- content: text response or audio file reference
- response_time: time to submission (seconds)
- timestamp: submission time

Logical Schema Representation

Figure 12 provides a conceptual representation of the logical data schema used by the assessment platform. The schema illustrates the relationships between participants, questions, and responses, abstracted from implementation-specific details. All identifiers are anonymized, and no personally identifiable information is stored within the research dataset.

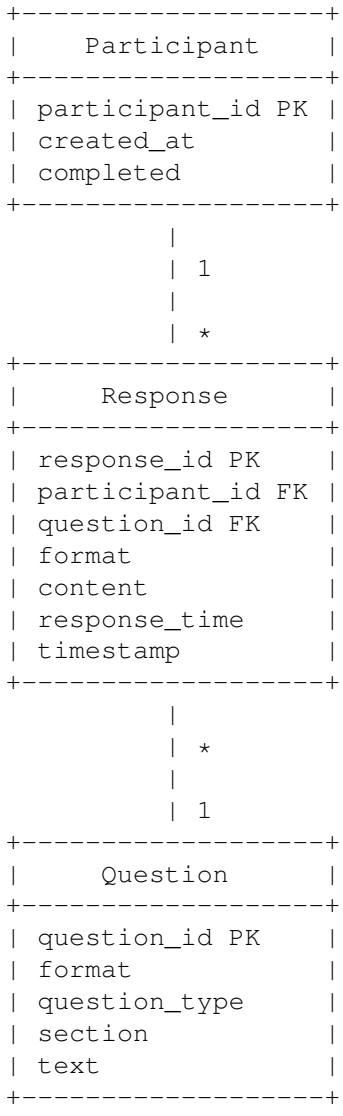

Fig. 12. Logical entity–relationship schema of the assessment platform database.

Derived metrics (accuracy scores, control consistency, and aggregate response times) were computed during post-processing and were not stored persistently in the database.

All identifiers were randomly generated and non-reversible. Authentication credentials were managed separately from the research dataset, and response content was accessible only to authorized researchers.

APPENDIX D
DATA PROCESSING AND ANALYSIS PIPELINE

This appendix summarizes the data processing and statistical analysis pipeline used in the study. The description is structured to enable replication of the analysis under equivalent conditions using the data schema defined in Appendix C.

Figure 13 presents a schematic overview of the data processing and statistical analysis pipeline. The diagram summarizes the sequential stages from participant enrollment through data filtering, metric computation, and inferential analysis. This structured representation complements the textual description and clarifies the execution order of the analytical workflow.

APPENDIX E
REPRODUCIBILITY AND DATA AVAILABILITY

All assessment materials, experimental protocols, and data structures required to reproduce this study are fully specified in Appendices A–D. Due to privacy and ethical considerations, raw response data and audio recordings are not publicly released.

An anonymized dataset and analysis scripts may be made available upon reasonable request for academic replication purposes, subject to institutional approval.

EXPLANATORY NOTES

National University of Science and Technology Politehnica Bucharest
The oldest and most prestigious technical university in Romania, founded in 1818. It recently underwent a merger and rebranding (formerly University Politehnica of Bucharest).

Faculty of Automatic Control and Computers
A leading academic faculty in Romania specializing in Computer Science, Information Technology, and Systems Engineering.

ANOVA (Analysis of Variance)
A statistical method used to compare the means of three or more groups to understand if at least one group differs significantly from the others. In this study, a two-way repeated-measures ANOVA was used to analyze interaction effects.

Shapiro-Wilk test
A statistical test used to determine if a sample comes from a normally distributed population, deciding whether parametric or non-parametric tests should be applied.

Cohen's d
An effect size used to indicate the standardized difference

between two means, helping to understand the practical significance of a result beyond simple statistical significance.

Wilcoxon signed-rank test
A non-parametric statistical hypothesis test used to compare two related samples to assess whether their population mean ranks differ, used here when normality was violated.

IER (Insufficient Effort Responding)
A phenomenon where participants provide responses with minimal cognitive effort (e.g., clicking randomly or ignoring instructions), which can invalidate study results.

TypeScript
A strongly typed programming language that builds on JavaScript, adding static type definitions to improve code reliability and maintainability in large-scale projects.

React
An open-source JavaScript library developed by Meta for building user interfaces, specifically designed for creating dynamic single-page applications (SPAs).

Firebase Realtime Database
A cloud-hosted NoSQL database provided by Google that allows data to be stored and synchronized in real-time across all connected clients.

NoSQL
A non-tabular database type that stores data differently than relational tables. It is often used for real-time applications and big data due to its scalability.

Web Audio API
A high-level JavaScript interface for processing and synthesizing audio in web applications, allowing the browser to capture and handle microphone input directly.

Bloom's Taxonomy
A hierarchical model used to classify educational learning objectives into levels of complexity, from basic recall to higher-order skills like evaluation and creation.

Universal Design for Learning (UDL)
An educational framework that guides the development of flexible learning environments to accommodate individual learning differences and ensure accessibility.

Joughin's Six Dimensions
A theoretical framework established by Gordon Joughin in 1998 used to analyze the structure and effectiveness of oral assessments in higher education.

UNSTPB (National University of Science and Technology Politehnica Bucharest)
The oldest and most prestigious technical university in Romania, founded in 1818. It recently underwent a merger and rebranding (formerly University Politehnica of Bucharest).

API (Application Programming Interface)
A set of rules and protocols that allows different software applications to communicate with each other. In this paper, it refers specifically to the Web Audio API used for capturing student responses.

NoSQL (Not Only SQL)
A non-tabular database type that stores data differently than relational tables, utilized by Firebase for its flexibility and real-time data synchronization.

UI/UX (User Interface / User Experience)
UI refers to the visual and interactive elements of the platform, while UX focuses on the overall ease of use and efficiency of the assessment process for the students.

SPA (Single Page Application)
A web application that interacts with the user by dynamically rewriting the current web page rather than loading entire new pages from a server, providing a smoother transition between questions.

JSON (JavaScript Object Notation)
A lightweight data-interchange format used for storing and transmitting assessment data between the frontend and the Firebase backend.

HTTP (Hypertext Transfer Protocol)
The standard protocol for transmitting data over the web, used for all communications between the student's browser and the application's server.

SD (Standard Deviation)
A statistical measure that quantifies the amount of variation or dispersion in a set of scores, used in the Results section to describe the consistency of student performance.

Tab Threshold
A specific data-filtering metric used in this study to identify participants who switched browser tabs excessively during the exam, which may indicate a lack of focus or external help.

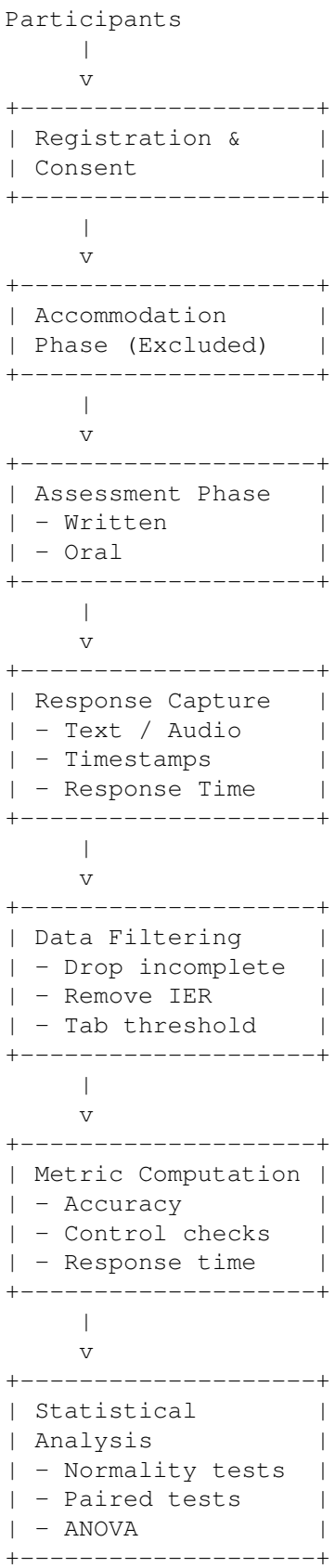

Fig. 13. Overview of the data processing and statistical analysis pipeline.