Jurnal Pendidikan MIPA
e-ISSN: 2685-5488 | p-ISSN: 1411-2531
URL: https://jpmipa.fkip.unila.ac.id/index.php/jpmipa
Jounal's DOI: http://doi.org/10.23960/jpmipa

Vol. 27, No. 3, 2026
Article ID: 1374
Pages: 1280-1297

Students' Learning Outcomes in Organic Chemistry Using Edpuzzle via Google Classroom

Dedi Irwandi[1,*], Pardiana[1], & R Ahmad Zaky El Islami[2]
[1]Department of Chemistry Education, Universitas Islam Negeri Syarif Hidayatullah Jakarta, Indonesia
[2]Faculty of Education, University of Pretoria, South Africa

*Corresponding email: dedi.irwandi.yuskar@uinjkt.ac.id

Received: 06 May 2026 *Accepted: 17 July 2026* *Published: 23 July 2026*

Abstract: This study aims to describe and analyze students' learning outcomes in the Organic Chemistry 1 course after the implementation of Edpuzzle-based interactive learning videos delivered through Google Classroom, and to examine the consistency of individual academic performance across two summative assessment points. A quantitative descriptive design was employed involving 52 undergraduate students enrolled in the Chemistry Education Program during the 2025/2026 Academic Year, consisting of Class A (n = 27) and Class B (n = 25). Four instructor-developed Edpuzzle videos (approximately 110 minutes in total) covering the topic of Atoms and Molecules were distributed via Google Classroom without embedded interactive checkpoints; the videos were designed for continuous, uninterrupted viewing. Learning outcomes were measured using essay-based Midterm (UTS) and Final (UAS) examinations, both of which were empirically validated and reliable (á = 0.801–0.882). Data were analyzed through descriptive statistics, the Shapiro-Wilk normality test, and correlation analysis (Pearson for Class A; Spearman for Class B). The results show that mean scores for Class A were 49.22 (UTS) and 43.52 (UAS), while Class B obtained 39.36 and 33.12, respectively, all below the institutional passing threshold. A strong positive UTS–UAS correlation was found in Class A ($r = 0.822$, $p < 0.001$) and Class B ($r_s = 0.788$, $p < 0.001$), indicating high relative rank-order stability of individual academic standing across the two assessment periods, rather than consistency of Organic Chemistry concept mastery (the two instruments differ in content domain and difficulty) or a causal effect of Edpuzzle. Interpreted through Mayer's Cognitive Theory of Multimedia Learning, the findings suggest that interactive video alone is insufficient to manage the intrinsic cognitive load of Organic Chemistry content. The study implies the need for structured video-based instructional designs integrating active learning, formative feedback loops, and early academic intervention for underperforming students.

Keywords: edpuzzle, video-based learning, learning outcomes, organic chemistry, cognitive theory of multimedia learning.

Article's DOI: https://doi.org/10.23960/jpmipa.v27i3.pp1280-1297

INTRODUCTION

The rapid digital transformation of higher education has changed the delivery of learning resources across many scientific disciplines. This is especially true for the field of Organic Chemistry. It is famous for being one of the most difficult subjects, as it involves abstract reaction mechanisms, complex molecular structures, and advanced techniques in problem solving (Galloway et al., 2018). Within this transformative context, learning platforms that utilize interactive video-based learning have immense potential as an active learning, conceptual visualization, and interactive engagement resource (Mayer, 2009; Giannakos et al., 2016).

As an interactive video learning system, EdPuzzle allows teachers to convert passive viewing to cognitive engagement by embedding formative questions, providing real-time feedback, and integrating real-time active viewing

(Mischel, 2019; Pulukuri & Abrams, 2020). When considering courses in which disengagement may be likely, the combination of interactive features, active engagement, and self-regulation may support heightened levels of attention and motivation (Cesare et al., 2021; Shelby & Fralish, 2021). For data-driven instructional design, combining EdPuzzle and Google Classroom (as an LMS) offers crucial options for monitoring real-time progress, providing personalized feedback, and assigning flexible tasks for both flipped and blended learning (Alim et al., 2019; Pulukuri & Abrams, 2020).

Video-integrated learning is centered on the Cognitive Theory of Multimedia Learning (Mayer, 2009). This theory states that when visual and verbal information are integrated and cognitive load is controlled, learners' comprehension is at an optimal level (Sweller et al., 2019). According to this framework, intrinsic cognitive load is inherently high in Organic Chemistry because learners must simultaneously process abstract symbolic representations (structural formulas), spatial-visual information (molecular geometry), and dynamic mechanistic processes (reaction pathways). When these element interactivity demands exceed working memory capacity, comprehension collapses regardless of the medium used (Skulmowski & Xu, 2022). This theoretical lens will be revisited in the Discussion section to interpret the observed learning outcomes. Empirically, Pulukuri & Abrams (2020) showed that integrating formative assessments within chemistry video modules improved students' accountability and understanding of concepts when combined with the flipped learning model. Improvements in laboratory experience and performance in biochemistry were also evident from the use of Edpuzzle (Shelby & Fralish, 2021). In the discipline of Organic Chemistry, Nsabayezu et al. (2025) show that integrating flipped learning with multimedia improves mastery of concepts and the ability to solve related problems. The use of Google Classroom along with Edpuzzle and other learning tools creates a data-driven, integrated instructional ecosystem within the context of cloud computing and learning (Alim et al., 2019).

Learning outcomes refer to students' cognition, attitude, and perception, which demonstrate achievement after the learning experience (Krathwohl, 2002). In the field of Organic Chemistry, many students have been observed performing poorly, owing to the abstract nature of the subject, the complexity of the concepts that require higher-order thinking, and low active engagement during the learning activity (Galloway et al., 2018; Nsabayezu et al., 2025). Although studies on Edpuzzle in chemistry education are growing, the majority have been conducted in developed-country contexts where digital infrastructure, prior exposure to self-regulated online learning, and language of instruction are qualitatively different from Indonesian higher-education contexts (Nsabayezu et al., 2025). Beyond mere geographic novelty, Indonesian chemistry undergraduates are situated within a broader contextual landscape that differs from developed-country settings in at least two respects: (i) infrastructural asymmetry, where variability in bandwidth, device ownership, and stable classroom connectivity is widely documented as a structural constraint on digital-learning engagement across Indonesian higher education generally (Rahiem, 2020; Weng & Wirda, 2025); and (ii) pedagogical-cultural factors, where teacher-centered didactic traditions have been argued in prior literature to reduce students' readiness for self-regulated video-based learning (Hairida et al., 2023). This premise is drawn from the broader literature on Indonesian higher education rather than tested directly in the present study, since the instrumentation described in the Method section was designed to capture cognitive learning outcomes rather than self-regulation, learning-readiness, or participants' pedagogical history. Because the present study recruited participants who met a baseline threshold of device and connectivity access, infrastructural access was deliberately controlled as a potential extraneous variable through the sampling procedure rather than examined empirically (Fraenkel et al., 2012; Creswell, 2014). This methodological choice is presented here as contextual background motivating why findings from infrastructurally distinct, developed-country contexts should not be assumed to generalize to Indonesian settings (Rahiem, 2020; Weng & Wirda, 2025), while the pedagogical-cultural dimension remains directly pertinent to interpreting the present results.

Based on the above research gap, this study aims to describe students' learning outcomes, specifically the Midterm (UTS) and Final (UAS) assessment scores in the Organic Chemistry 1 course after Edpuzzle-based instructional materials were delivered through Google Classroom in the 2025/2026 academic year. Accordingly, this study is guided by three research questions: (RQ1) How are the UTS and UAS scores of Class A and Class B distributed after the Edpuzzle-based learning intervention? (RQ2) How can the observed patterns of scores be interpreted with reference to the cognitive demands of Organic Chemistry? (RQ3) To what extent does individual academic performance remain consistent between UTS and UAS, as reflected in the UTS–UAS correlation within each class?

METHOD

Research Design and Procedures

This study used a quantitative descriptive approach (Creswell, 2014) to systematically and objectively analyze students' learning outcomes in the Organic Chemistry 1 course. A descriptive approach was intentionally chosen to analyze learning outcomes and achievements under real circumstances, as opposed to researching the causes of learning outcomes (Fraenkel et al., 2012). This aligns with Sukardi (2011), who argues that quantitative descriptive research is relevant for analyzing the patterns, trends, and characteristics of educational phenomena. Because the design is descriptive and non-experimental, a control group is neither required nor appropriate; the study does not aim to establish a causal effect of Edpuzzle, but rather to describe learning-outcome patterns and their consistency in an authentic instructional setting. This design choice, however, constrains the interpretability of absolute performance indicators such as the sub-threshold mean scores reported in Table 3: without a comparison group taught through conventional instruction, or normative data from prior offerings of the course using equivalent instruments, it is not possible to determine whether these scores primarily reflect a characteristic of the Edpuzzle-based intervention, the inherent difficulty of the examination instruments, or both. This interpretive constraint is acknowledged here and addressed explicitly, rather than resolved, in the Limitations subsection. The study was conducted over one full semester (16 academic weeks) in the 2025/2026 academic year, following the sequence: (1) development and pilot review of four instructor-produced Edpuzzle videos; (2) distribution of the videos through Google Classroom to both classes during Weeks 3–7; (3) administration of the Midterm Examination (UTS) in Week 8; (4) continued instruction and reinforcement activities in Weeks 9–15; and (5) administration of the Final Examination (UAS) in Week 16.

Participants

The study sample included 52 students from the Chemistry Education Program who enrolled in Organic Chemistry 1 during the 2025/2026 academic year, which consisted of Class A (n =

27) and Class B (n = 25). A purposive sampling technique was applied. The eligibility criteria for participation were as follows: (a) formally registered in the Organic Chemistry 1 course during the 2025/2026 academic year at the study institution; (b) had not previously taken Organic Chemistry 1 (i.e., first-time enrollees, to avoid the confounding effect of prior exposure); (c) had regular access to a personal device (smartphone or laptop) with sufficient internet connectivity to access Edpuzzle and Google Classroom; (d) completed both the Midterm (UTS) and Final (UAS) examinations; and (e) provided informed consent for their examination scores to be used for research purposes. Criterion (c) was applied to ensure that all participants could engage with the Edpuzzle-based instructional materials as intended; consequently, infrastructural access was deliberately held constant across the sample, and this study does not empirically test infrastructural asymmetry as a variable but instead isolates cognitive and pedagogical factors under conditions of adequate digital access. This methodological choice, and its implications for the infrastructural rationale raised in the Introduction, are revisited in the Limitations subsection. Although Organic Chemistry 1 institutionally requires prior completion of Basic Chemistry I and Basic Chemistry II, and all participants therefore satisfied this prerequisite by virtue of formal enrollment, quantitative baseline indicators of participants' prior academic standing such as their cumulative grade point index (IPK) from the preceding semester or their grades in these prerequisite courses were not collected. The cognitive equivalence of Class A and Class B at the outset of the semester could therefore not be empirically verified, and the descriptive comparison between the two classes reported in Table 3 should be interpreted with this constraint in mind. Students who withdrew from the course, transferred classes mid-semester, or missed either assessment were excluded from the analysis.

Instruments

The measurement tools used in this research were two institutionally administered essay-based tests, the Midterm Examination (UTS) and the Final Semester Examination (UAS), each scored on a 0–100 scale. The items were constructed by the teaching team based on the Course Learning Outcomes (CPMK) established by the study program, covering the cognitive domains from understanding (C2) to analyzing (C4) of the revised Bloom's Taxonomy (Krathwohl, 2002). The UTS comprised 16 items for both Class A and Class B, while the UAS comprised 13 items for both classes.

Scoring System and Inter-Rater Reliability

Each essay item was scored analytically by decomposing it into sub-components with pre-established point weights (maximum raw scores: UTS = 127, UAS 1 = 129, UAS 2 = 126). The raw scores were then summed and converted to a 0–100 scale using the formula: Score (0-100) = (total raw score / maximum raw score) × 100. This approach was chosen because organic chemistry essay responses contain clearly right-or-wrong components such as definitions, structures, reaction mechanisms, and nomenclature that lend themselves to explicit, criterion-referenced scoring. To reduce rater bias, a double-anonymized grading procedure was implemented. Every answer sheet was scored independently by two raters: the course instructor (Rater 1) and an independent second rater (Rater 2). Both raters evaluated the responses without knowledge of student identities or each other's scores, utilizing an identical analytic rubric that operationalized the original scoring guideline into explicit, item-level criteria.

The reported score for each student was the average of the two raters' scores. Inter-rater consistency was evaluated using the Pearson correlation and the mean absolute score difference on the 0–100 scale across all 52 students in both classes. Agreement was excellent for all instruments (UTS: r = 0.980; UAS 1: r = 0.988; UAS 2: r = 0.976; mean absolute difference < 5 points), exceeding the r e" 0.75 threshold for "excellent agreement" (Cicchetti, 1994) and indicating low inter-rater bias following the double-anonymized grading procedure.

Test blueprint

To document the depth and coverage of the cognitive domains being assessed, a test blueprint mapping each item to its intended cognitive level was developed prior to instrument administration (see Table 1).

To substantiate the C4 (Analyze) classification assigned to Sub-CPMK 04.5.4, a representative item from the UTS instrument was reproduced below rather than relying solely on the action verb stated in the syllabus. The item required students to determine the R/S configuration of five distinct chiral compounds: "*Tentukan R atau S senyawa berikut dengan urutan, tulis atom yang terikat pada C kiral, urutkan prioritas, proyeksikan prioritas terkecil menjauhi pengamat*" ("Determine the R or S configuration of the following compounds in sequence; identify the atoms bonded to the chiral carbon, rank the substituent priorities, and project the lowest-priority group away from the observer"). Answering this item requires students to (a) identify the chiral center within each structure, (b) decompose and rank substituents according to Cahn-Ingold-Prelog priority rules, (c) mentally reorganize the three-dimensional spatial arrangement into the correct Newman-type projection, and (d) integrate these sub-operations to derive the final configurational assignment. This multi-step decomposition and spatial reorganization of parts is consistent with the C4 (Analyze) category of the revised Bloom's Taxonomy (Krathwohl, 2002), rather than the or single-step application characteristic of C1-C3.

Table 1. Cognitive blueprint of the UTS and UAS instruments (mapped to course Sub-CPMK)

Sub-CPMK	Learning Outcome	Bloom's Level	Instrument Allocation
04.5.2	Explaining the concepts of atoms and molecules	C2 (Understand)	UTS
04.5.3	Explaining orbitals and their role in covalent bonding	C2 (Understand)	UTS
04.5.4	Analyzing the stereochemistry of a compound	C4 (Analyze)	UTS
04.5.5	Explaining types, properties, nomenclature, and reactions of alkyl halides	C2 (Understand)	UAS
04.5.6	Analyzing free radical reactions	C4 (Analyze)	UAS
04.5.7	Explaining types, properties, nomenclature, and reactions of alcohols and ethers	C2 (Understand)	UAS
04.5.8	Explaining types, properties, nomenclature, and reactions of alkenes and alkynes	C2 (Understand)	UAS

Note. The blueprint was constructed by mapping each examination item to the Sub-CPMK defined in the course syllabus. Bloom's cognitive level is inferred from the action verb of each Sub-CPMK (Krathwohl, 2002).

Internal consistency reliability was estimated using Cronbach's á, with a minimum threshold of 0.70 for educational measurements (Hair et al., 2019). All four instruments yielded á values above this threshold: 0.863 (UTS Class A), 0.873 (UTS Class B), 0.882 (UAS Class A), and 0.801 (UAS Class B), all of which are classified as acceptable to good reliability (George & Mallery, 2003).

Content delivery of the Organic Chemistry 1 course is achieved using instructor-created videos placed on Edpuzzle and given out via Google Classroom. For students, the requirements are to (1) watch the content of videos, (2) listen to the content and take notes, without any embedded quiz or checkpoint response required. This is done for both classes and is adhered to for the whole semester.

The learning intervention of this study used Edpuzzle as a learning tool on the topic of Atoms and Molecules in the Organic Chemistry 1 course. A total of four videos, totaling about 110 minutes, were created and uploaded by the instructor on Edpuzzle. These videos cover the following sub-topics: (1) Structural Formulas and Spatial Structural Formulas (17 minutes 55 seconds), (2) Ionic Bonds, Covalent Bonds, and Formal Charge (33 minutes 16 seconds), (3) Electron Structure, Atomic Radius, and Electronegativity (29 minutes 18 seconds), and (4) Bond Lengths, Angles, and Polar Covalent Bonds (30 minutes 14 seconds). The videos did not embed any interactive questions or checkpoints; playback proceeded continuously without pauses requiring a student response. Because Edpuzzle assignments are viewable to students in each class according to the timetable, it supports a flexible, regular style of learning.

Course content was delivered through instructor-created videos placed on Edpuzzle and distributed via Google Classroom. Four videos were developed by the instructor (Dedi Irwandi) with a total duration of approximately 110 minutes, covering (1) Structural Formulas and Spatial Structural Formulas (17 min 55 s); (2) Ionic Bonds, Covalent Bonds, and Formal Charge (33 min 16 s); (3) Electron Structure, Atomic Radius, and Electronegativity (29 min 18 s); and (4) Bond Lengths, Angles, and Polar Covalent Bonds (30 min 14 s). The videos did not embed any interactive checkpoint questions, quizzes, or response prompts; playback was continuous and non-interruptible, requiring only passive viewing and listening. The pedagogical design of the four videos was aligned with one of Mayer's Cognitive Theory of Multimedia Learning principles: coactivation of verbal and pictorial channels; narration was synchronized with animated 2D/3D molecular representations. Notably, the videos in this study did not provide any embedded formative questions, automated correctness feedback, adaptive scaffolding, branching prompts, worked-example fading, or post-video collaborative activity. This absence of interactive and formative-feedback elements represents an intentional baseline whose limitations are revisited in the Discussion.

Table 2. Summary of empirical validity and reliability of UTS and UAS instruments

Instrument	Class	No. of Items	Valid Items	α Cronbach
UTS	A	16	16 (100%)	0.863
UTS	B	16	16 (100%)	0.873
UAS	A	13	13 (100%)	0.882
UAS	B	13	13 (100%)	0.801

Note. α = Cronbach's alpha; UTS = Midterm Examination; UAS = Final Examination

It should be noted that the recorded durations (17 min 55 s to 33 min 16 s) exceed the short-segment chunking commonly associated with the segmentation principle within Mayer's

Cognitive Theory of Multimedia Learning (Mayer, 2009). This departure was not incidental: as stated above, the four videos were designed to instantiate only one Mayer principle, namely dual-channel coactivation of narration and animation, while segmentation through embedded checkpoints was deliberately left inactive so that each video could be viewed as a single, uninterrupted, lecture-equivalent unit corresponding to one coherent conceptual topic (e.g., Ionic Bonds, Covalent Bonds, and Formal Charge). Consequently, the intrinsic cognitive load associated with Organic Chemistry content in this study was not offset by any structural chunking mechanism, and the resulting implication for working-memory management is examined further in the Discussion section.

Data Analysis

The three-step data analysis included: (1) descriptive statistics documenting the mean, standard deviation (SD), median, minimum and maximum values, and frequency distributions of value ranges; (2) the Shapiro-Wilk normality test to identify correlation methods; and (3) analysis of the correlations between UAS and UTS scores; and (4) a post-hoc statistical power analysis for each reported correlation coefficient, undertaken specifically to evaluate whether the per-class sample sizes (N = 25–27) afforded adequate sensitivity to detect the observed effects, given that correlation analyses based on fewer than 30 cases per group are, in principle, more susceptible to sampling variability. For UAS and UTS scores, class performance in assessment A (normal distribution) was analyzed using the Pearson correlation. In contrast, class performance in assessment B (non-normal distribution) was analyzed using the Spearman correlation (Field, 2018).

Prior to correlation analysis, boxplot inspection using the 1.5 × IQR criterion identified two Class B scores as statistical outliers (UTS = 94; UAS = 69; see Figure 1). Both values were verified as legitimate, non-erroneous examination results and were therefore retained in the dataset rather than removed, winsorized, or otherwise adjusted. This retention was considered appropriate because Class B was subsequently analyzed using the Spearman rank correlation, which operates on rank-transformed rather than raw values and is thus inherently less sensitive to extreme scores than the Pearson correlation applied to Class A (Field, 2018). Correlations were interpreted as: very weak (< 0.20), weak (0.20–0.39), moderate (0.40–0.59), strong (0.60–0.79), and very strong (0.80–1.00) following Evans (1996).

To evaluate whether the per-class sample sizes (N = 25–27) were adequate for producing stable and reliable population correlation estimates, a post-hoc power analysis was performed for each reported coefficient. This analysis used the Fisher r-to-z transformation method for correlation power estimation (Cohen, 1988), which aligns with the correlation-power routine in G*Power 3 (Faul et al., 2007). For each class, the observed correlation coefficient and sample size were used to compute the achieved power (1 “ â) of the two-tailed significance test at á = 0.05, together with the 95% confidence interval for the coefficient derived from the same Fisher z-transformation. All analyses used á = 0.05. Since the UTS and UAS instruments assessed different content areas with varying item difficulty levels, direct comparisons of their mean scores should not be seen as evidence of academic decline. Instead, they were reported descriptively and interpreted cautiously. The heterogeneity of these instruments also limits the interpretation of the UTS-UAS correlation shown in Table 4. Because the two assessments focus on different content domains and difficulty levels (see Table 1), the correlation coefficient is understood only as an indicator of the relative rank-order stability of individual students within each class, not as proof of consistent mastery of Organic Chemistry concepts due to the intervention.

RESULT AND DISCUSSION

Distribution of UTS and UAS Scores (RQ1)

The descriptive statistics for the UTS and UAS scores in Class A and Class B are presented in Table 3.

Table 3. Descriptive statistics and normality test results (shapiro-wilk) of UTS and UAS scores for class A and class B

Class	Test	N	Mean	SD	Median	Min	Max	SW W	SW p	Distribution
Class A	UTS	27	49.22	19.18	53.00	9	92	0.987	.976	Normal
Class A	UAS	27	43.52	21.49	41.00	6	86	0.977	.787	Normal
Class B	UTS	25	39.36	17.46	36.00	17	94	0.906	.024*	Non-normal
Class B	UAS	25	33.12	13.00	32.00	16	69	0.913	.036*	Non-normal

Note. SW = Shapiro–Wilk normality test (df = N). *p < .05 indicates deviation from normality. The Shapiro–Wilk test was reported because it is more appropriate than the Kolmogorov–Smirnov test for small samples (N < 50), which characterized all groups (N = 25–27) in this study.

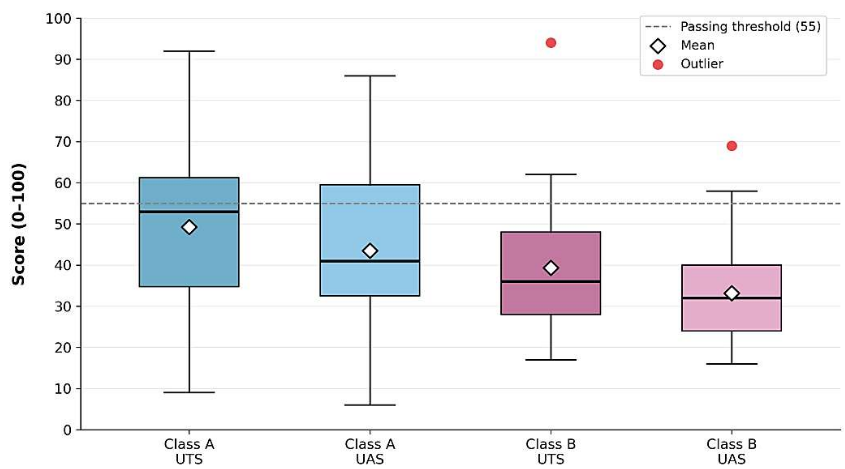

Figure 1. Distribution of UTS and UAS scores by class (side-by-side boxplot). Note: Boxes show the interquartile range (Q1–Q3); solid horizontal lines within the boxes show the median; diamond markers show the mean; whiskers extend to the most extreme non-outlier data points; red circles indicate outliers (values beyond 1.5 × IQR from the nearest quartile). The dashed horizontal line marks the institutional passing threshold at 55.

Although the boxplot in Figure 1 summarizes the distribution of scores at each assessment point, it does not, by design, show whether the same students occupied similar relative positions from UTS to UAS. To convey this individual-level information directly, and following the recommendation that paired or repeated-measures data be visualized so that each participant's trajectory remains traceable rather than collapsed into group summary statistics (Weissgerber et al., 2015), a paired slope graph was constructed by connecting each student's UTS and UAS scores with a line segment (Figure 2).

Figure 2 shows that the downward shift documented in Figure 1A and Table 3 was not uniform across students but reflected a majority pattern with a substantial minority moving in the

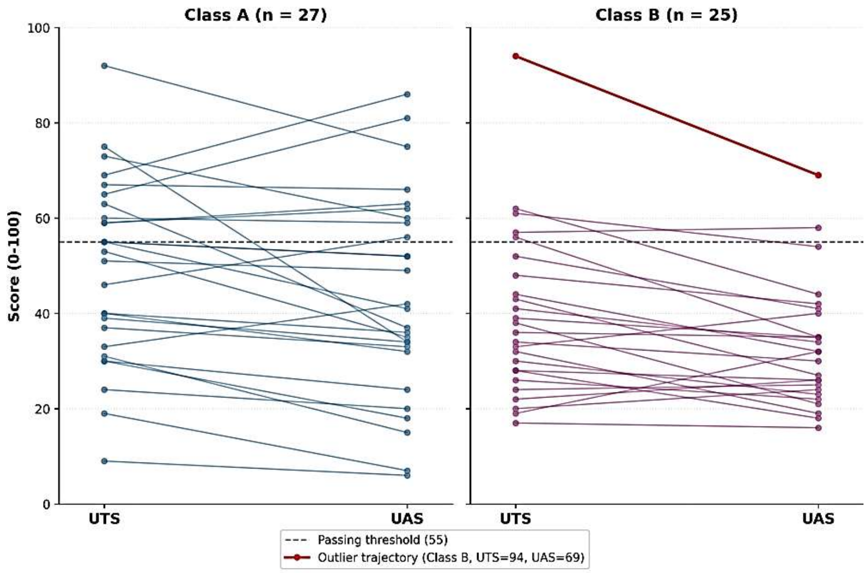

Figure 2. Paired slope graph of individual UTS-to-UAS score trajectories, by class. Note. Each line connects one student's UTS score to their UAS score; the dashed horizontal line marks the institutional passing threshold at 55. The red line highlights the trajectory of the Class B student identified as a statistical outlier in Figure 1A (UTS = 94, UAS = 69)

opposite direction. In Class A, 21 of 27 students (77.8%) scored lower on the UAS than on the UTS, while 6 (22.2%) improved; relative to the passing threshold of 55, 8 students remained above the threshold at both assessments, 5 fell from above to below, 1 rose from below to above, and 13 remained below the threshold throughout. In Class B, 19 of 25 students (76.0%) declined from UTS to UAS and 6 (24.0%) improved; 2 students remained above the threshold at both assessments, 3 fell from above to below, none rose from below to above, and 20 remained below the threshold throughout. The single Class B trajectory identified as an outlier in Figure 1A (UTS = 94, UAS = 69) is visibly the steepest decline in either class, illustrating at the individual level the same data point that the boxplot flags only as a distributional extreme.

Table 3 and Figure 1 together reveal three substantive patterns rather than isolated averages. First, both classes clustered well below the institutional passing threshold of 55: the median line for every group is below the reference line, and none of the four boxes has an upper quartile above 65. Second, Class A shows a wider interquartile range (IQR = 26.5 for UTS; 27.0 for UAS) than Class B (IQR = 20.0 for UTS; 16.0 for UAS), indicating substantially more within-class heterogeneity in Class A. Third, Class B's UAS distribution compresses toward the lower end (IQR = 16.0; whisker minimum = 16), signaling that many Class B students plateaued at a low performance level rather than reflecting improved consistency of high achievement; a single Class B student (score = 69) appears as an outlier above the main distribution. These observations align with Nsabayezu et al. (2025), who reported that students continue to struggle with abstract organic chemistry concepts even when multimedia is used, particularly when the instructional design does not explicitly manage element interactivity.

The majority of students in both classes placed in the lowest grading category (E, <40). Because different test blueprints and item difficulties preclude a direct comparison of UTS and UAS means as evidence of academic decline (Field, 2018), the distributional shift is interpreted with caution. Descriptively, however, the proportion of students in the lowest band moved from 33.3% to 48.1% in Class A and from 60.0% to 72.0% in Class B, and no student in either class reached the highest achievement band on the UAS. This descriptive pattern is consistent with the observation by Pulukuri and Abrams (2020) that, without clear course structuring and clear learning incentives, students' sustained engagement with Edpuzzle-based instruction tends to erode over time.

Figure 3 visualizes the proportional shift more directly than the raw counts. In Class A, the combined below-passing band (D + E) expanded from 51.8% at UTS to 66.6% at UAS. In Class B, the same combined band expanded from 80.0% to 92.0%. The passing-or-above band (A + B + C) therefore contracted by roughly 15 percentage points in Class A and 12 percentage points in Class B between the two assessments, with the concentration into the E band being especially pronounced in Class B.

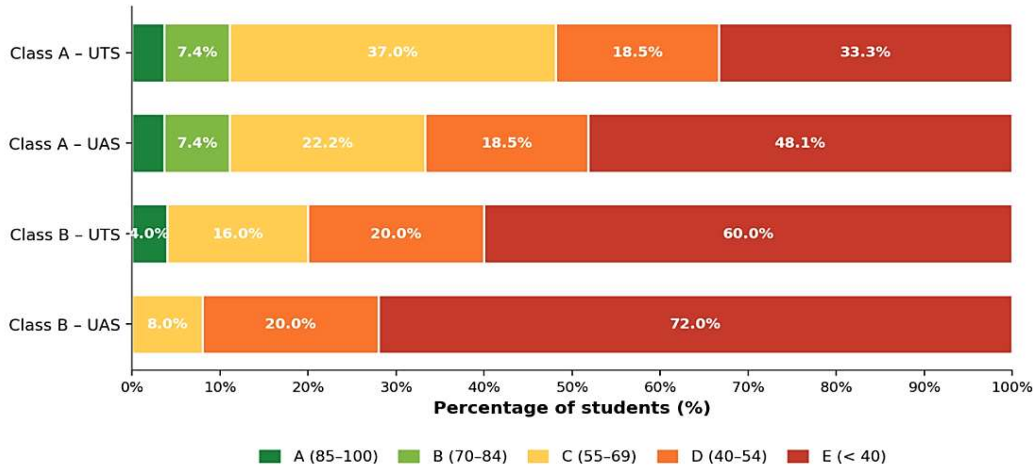

Figure 3. Proportional distribution of grade categories per class and assessment (100% stacked bar chart). Note: A = 85–100 (Excellent); B = 70–84 (Good); C = 55–69 (Passing); D = 40–54 (Below Passing); E < 40 (Fail). Percentages are computed within each class-assessment combination

Interpretation through the Lens of Cognitive Load and Content Demand (RQ2)

The persistently low performance observed in both classes cannot be definitively attributed to a single cause, as cognitive load was not directly measured in this study. Because the study did not include a group taught through conventional instruction, the observed sub-threshold performance may equally reflect the intrinsic stringency of the essay-based UTS and UAS instruments, particularly the C4-level items requiring multi-step spatial reasoning (Table 1), independent of the instructional method used. The validity and reliability evidence reported in Table 2 concerns internal consistency and item-total correlation, not absolute item difficulty, and therefore does not resolve this ambiguity. Disentangling the relative contribution of instructional design and instrument difficulty to the observed performance would require a comparison condition that this descriptive design does not provide.

The distributional shift toward the lower category from UTS to UAS, while not causally attributable to Edpuzzle use, is consistent with two documented phenomena: increased element interactivity in Organic Chemistry content between the earlier topics (Atoms and Molecules) and the later topics (reaction mechanisms, functional group transformations) that require higher-order thinking (Galloway et al., 2018; Nsabayezu et al., 2025), and end-of-semester academic fatigue that reduces sustained self-regulated engagement with asynchronous video content (Okoth, 2025). Frequent formative assessments across the semester rather than reliance on end-of-topic quizzes have been recommended to sustain motivation and provide timely correction of misconceptions (Okoth, 2025; Cornelison et al., 2022). It must be emphasized that this comparison of means is descriptive only, as UTS and UAS assessed different topics with different item difficulties, precluding statistical inference (Field, 2018).

Consistency of Individual Academic Performance: The UTS–UAS Correlation (RQ3)

Prior to correlation analysis, the Shapiro-Wilk normality test was conducted for each class and each assessment, as it is more appropriate than the Kolmogorov-Smirnov test for the small sample sizes ($N < 50$) in this study. Results are summarized together with the descriptive statistics in Table 3 above.

As indicated in Table 3, under the Shapiro-Wilk criterion, Class A data satisfied the normality assumption (UTS: $p = 0.976$; UAS: $p = 0.787$), whereas Class B did not (UTS: $p = 0.024$; UAS: $p = 0.036$). This pattern is consistent with the distributional shape reported in Table 3: Class A exhibited near-symmetric distributions with skewness close to zero, whereas Class B exhibited substantial positive skewness (> 1) and positive excess kurtosis, indicating right-skewed distributions with a heavy lower tail. Accordingly, the Pearson correlation was applied to Class A and the Spearman rank-order correlation to Class B. The results are reported in Table 4.

Figure 4 visualizes the correlation reported in Table 4 at the individual-student level. In both classes, the majority of points fall below the $y = x$ reference line, indicating that most students obtained a lower UAS than UTS score, consistent with the descriptive downward shift discussed in Sub-section RQ1. The clustering of points along rather than randomly around the diagonal confirms the strong rank stability reflected in the correlation coefficients. This combined visualization also reveals that both classes contain a small number of students who performed above the passing threshold on both assessments (top-right quadrant), and a considerably larger group who remained below the threshold on both (bottom-left quadrant), reinforcing the pattern of rank-order stability already reflected in the correlation coefficients reported in Table 4.

Table 4. UTS–UAS correlation results by class

Class	Method	Coefficient	N	p	95% CI	Post-hoc Power (1−β)	Interpretation
Class A	Pearson	$r = 0.822$	27	< 0.001	[0.643, 0.916]	0.999	Very strong, positive
Class B	Spearman	r_s = 0.788	25	< 0.001	[0.571, 0.902]	0.999	Strong, positive

Note. Interpretation follows Evans (1996): 0.60–0.79 = strong; 0.80–1.00 = very strong. As the UTS and UAS instruments differ in content domain and item difficulty (Table 1), these coefficients index only the relative rank-order stability of individual students within each class and do not indicate consistency of concept mastery attributable to the intervention. 95% CI and post-hoc power (1 “ â) were computed via the Fisher r-to-z transformation (Cohen, 1988) for a two-tailed test at á = 0.05.

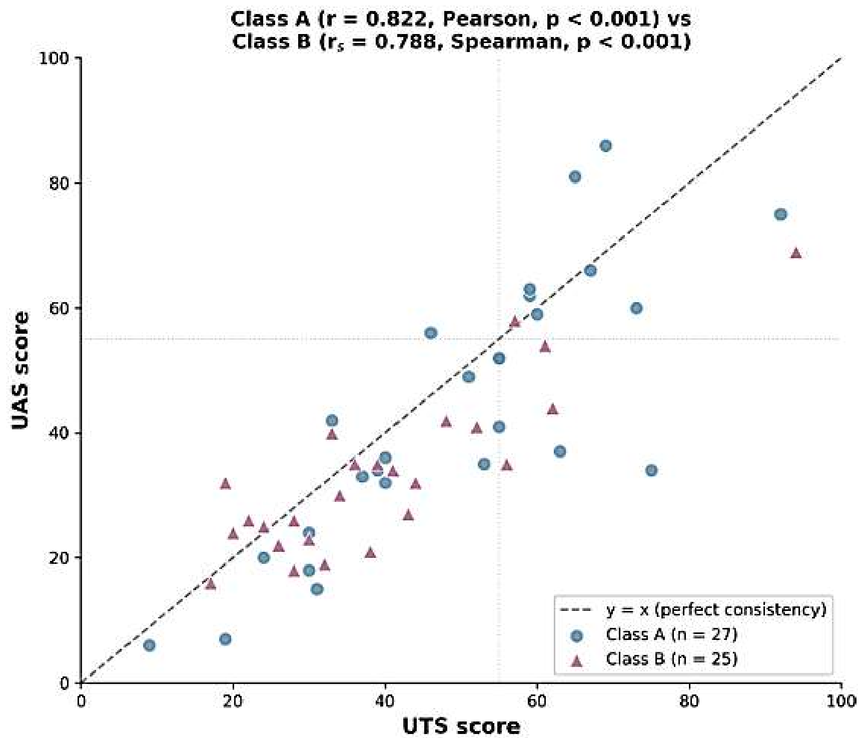

Figure 4. Combined scatter plot of UTS × UAS scores for Class A and Class B, distinguished by marker color and shape (circles for Class A, triangles for Class B), with a y = x diagonal reference line

Both classes exhibited a strong positive correlation between UTS and UAS scores (Class A: r = 0.822, p < 0.001; Class B: r_s = 0.788, p < 0.001). It is critical to interpret these findings within their methodological limits. Because the UTS and UAS instruments assessed different content domains with different levels of item difficulty (Table 1), this coefficient must not be read as evidence of consistent mastery of Organic Chemistry concepts resulting from the intervention; its scope is restricted to the relative rank-order stability of individual students' standing within each class. A further methodological concern, raised during the review process, is that correlation analyses based on fewer than 30 cases per group are, in principle, more susceptible to sampling variability. This concern was addressed directly through the post-hoc power analysis reported in Table 4: given the observed effect sizes (r = 0.822 for Class A; r_s = 0.788 for Class B), the achieved statistical power for detecting these correlations at á = 0.05 (two-tailed) was .999 for both classes, well above the conventionally recommended minimum of .80 (Cohen, 1988). The corresponding 95% confidence intervals ([0.643, 0.916] for Class A; [0.571, 0.902] for Class B) are narrow relative to the 0–1 range of the coefficient and exclude weak or negligible values entirely.

Taken together, these results indicate that although the per-class sample sizes are modest in absolute terms, the very large magnitude of the observed associations provided ample sensitivity for the significance tests reported. Hence, the reliability of the reported coefficients is not meaningfully compromised by the sample size alone. The correlation coefficients quantify the consistency of individual academic rank between two assessment points; they do not, and cannot, provide causal evidence of Edpuzzle's effectiveness because the study lacks a counterfactual comparison group. In other words, a student who performed poorly at midterm tended to perform poorly at the final, and a student who performed well tended to remain among the higher performers. However, this stability may reflect prior knowledge, self-regulation capacity, or general study habits rather than any effect of the interactive video treatment.

The pedagogical implication, however, remains meaningful. The high rank-order stability suggests that students who receive a D or E grade at midterm are unlikely to substantially improve their standing by the final exam without additional support, which points to midterm results as a practical trigger for instructor attention. Establishing UTS performance as a formal early-warning indicator would nonetheless require finer-grained longitudinal data, such as trends in weekly formative-quiz scores, which were not collected in this study. Whether the Edpuzzle analytics dashboard (e.g., rewatch frequency, completion time) can likewise be developed into a validated early-warning tool, as proposed by Shelby and Fralish (2021) and Zhang et al. (2006), remains a direction for future research rather than a claim substantiated by the present data.

Implications for Instructional Design

The overall low performance across both classes highlights that interactive videos alone are unlikely to be sufficient for teaching Organic Chemistry. Although Edpuzzle provides interactive tools that support engagement (Mischel, 2019; Pulukuri & Abrams, 2020), the effectiveness of such tools depends on the surrounding instructional design. As documented in the Method section, the four videos used in this study incorporated only dual-channel coactivation of narration and animation, without embedded formative questions, interactive checkpoints, adaptive branching, worked-example fading, or post-video collaborative discussion. These absent design elements are known correlates of successful multimedia learning in high-load domains (Van Gog et al., 2020); their absence is therefore a plausible explanation for the observed ceiling on learning outcomes, and should be treated as a design limitation rather than as an empirical claim about scaffolding effects, since scaffolding was not directly measured in this study. This interpretation is anchored in the data reported above rather than in external literature alone: the cognitive blueprint in Table 1 identifies Sub-CPMK 04.5.4 (analyzing the stereochemistry of a compound, C4) as the sole Analyze-level item on the UTS, and the representative R/S-configuration item reproduced in the Instruments section requires students to integrate chiral-center identification, Cahn-Ingold-Prelog priority ranking, and three-dimensional spatial projection within a single multi-step operation. This is precisely the class of high-element-interactivity task for which the absence of worked-example fading and checkpoint-based formative feedback would be expected to matter most. The descriptive shift toward the lowest achievement band (Table 3; Figure 2)with the proportion of students in the E band rising from 33.3% to 48.1% in Class A and from 60.0% to 72.0% in Class B, and no student reaching the highest band on the UASis consistent with, though it does not by itself demonstrate, a concentration of difficulty on this type of C4-level item.

Based on studies of multimedia-assisted flipped learning in Organic Chemistry (Nsabayezu et al., 2025), video-based learning produces the best results when combined with active learning strategies such as collaborative problem-solving, case-based discussions, and structured peer interactions regularly. Research in pharmacy education has similarly documented that interactive video modules are most effective when combined with structured follow-up activities and ongoing formative assessments

(Cornelison et al., 2022). It should be noted that the present study did not conduct an item-level or sub-component error analysis of students' UTS/UAS answer sheets: as described in the Instruments section, sub-component points for each essay item were aggregated into a single 0–100 score per instrument, and no record was kept of which specific reasoning step (e.g., chiral-center identification, CIP priority ranking, or three-dimensional projection) individual students failed on a given item. The design recommendations below are therefore derived from the cognitive-level evidence the study did collect, the Table 1 blueprint and the worked C4 item example together with the aggregate score-band shift reported in Table 3 and Figure 2, rather than from a diagnosed pattern of specific student errors; this constitutes an additional limitation of the present design, addressed further below.

Extending these recommendations to the present setting, three concrete design refinements are proposed for subsequent iterations of Edpuzzle-based Organic Chemistry instruction: (a) inserting worked-example steps with fading immediately before high-load conceptual segments, such as the stereochemical-analysis segment corresponding to Sub-CPMK 04.5.4 (Table 1), which the R/S-configuration item in the Instruments section illustrates as the most cognitively demanding component of the UTS; (b) coupling each Edpuzzle module with a mandatory post-video collaborative task; and (c) using the Edpuzzle analytics dashboard to trigger targeted micro-tutorials for students whose video engagement metrics (e.g., rewatch frequency, completion time) fall below a defined threshold, prioritizing students who scored within the lowest achievement bands at UTS (Table 3), consistent with the early-warning rationale discussed under RQ3. Future iterations of this design should be evaluated using item- or sub-component-level error analysis of student responses, for example, coding which CIP-ranking or projection sub-steps are most frequently missed so that subsequent design recommendations can be derived directly from documented student error patterns rather than from cognitive-level item classification and aggregate score-band shifts alone.

■ CONCLUSION

This study set out to describe students' learning outcomes in an Edpuzzle-supported Organic Chemistry 1 course and to examine the consistency of individual academic performance across two summative assessments. Answering RQ1 and RQ2, the score distributions in both classes clustered below the institutional passing threshold, with the majority of students concentrated in the lowest achievement band, a finding that, given the descriptive design and the absence of a direct measure of cognitive load, should be interpreted as consistent with, rather than proof of, the instructional design limitations and instrument stringency discussed above. Answering RQ3, the strong positive UTS–UAS correlations in both classes indicate a high stability of individual academic rank across the semester rather than a causal effect of Edpuzzle; given that the two instruments differ in content domain and difficulty (Table 1), this correlation does not indicate consistency of Organic Chemistry concept mastery; midterm and final performance track each other closely at the individual level across the semester. Taken together, these findings reposition Edpuzzle-with-Google-Classroom not as a stand-alone solution for high-load conceptual content, but as one component of a broader, load-managed instructional ecosystem that must be paired with active-learning, formative-feedback, and early-intervention layers to yield meaningful gains.

The implications for practice and research follow directly. For instructional practice, chemistry lecturers using Edpuzzle should treat the platform's analytics as a formative monitoring tool rather than as evidence of learning. They should embed worked-example fading, structured peer discussion, and formative-feedback cycles around the video content to manage the high intrinsic load of Organic Chemistry. At the institutional level, instructor professional development in multimedia instructional design and the availability of adequate learning-support infrastructure remain preconditions for realizing the pedagogical potential of interactive video. This study also has clear limitations that circumscribe its conclusions. Its descriptive design precludes causal inference, and its single-institution, single-cohort sample limits generalizability. More specifically, the absence of a comparison group taught through conventional instruction means that the sub-threshold mean scores reported in Table 3 cannot be unambiguously attributed to limitations of the Edpuzzle-based intervention. Although the UTS and UAS instruments underwent construct-validity and reliability testing (Table 2), this evidence concerns internal consistency and item-total correlation rather than absolute item difficulty; it therefore remains possible that the essay instruments, particularly the C4-level items, were demanding independent of the instructional method used. Distinguishing an instructional-design explanation from an instrument-difficulty explanation for the observed low scores would require either a conventionally taught comparison group or historical benchmarking data from prior cohorts assessed with equivalent instruments, neither of which was available in this study. The quasi-experimental comparison design recommended below would allow these two competing explanations to be empirically disentangled. Relatedly, because no baseline measure of participants' prior academic standing, such as prior-semester cumulative GPA or grades in the prerequisite courses Basic Chemistry I and Basic Chemistry II, was collected, the cognitive equivalence of Class A and Class B at the start of the semester could not be empirically verified, further limiting the interpretability of the between-class comparison presented in Table 3.

To reduce rater bias, a double-anonymized grading procedure was implemented. Every answer sheet was scored independently by two raters. Inter-rater agreement was excellent across instruments ($r = 0.976–0.988$), supporting the objectivity of the scoring procedure. No digital-infrastructure instrument was administered, so environmental factors could not be empirically tested. Similarly, no instrument measuring self-regulated learning readiness, self-regulation capacity, or participants' prior pedagogical exposure was administered in this study. Consequently, the pedagogical-cultural rationale invoked in the Introduction functions as contextual background motivating the study rather than as an empirically tested explanatory variable, and should not be interpreted as a demonstrated cause of the observed learning outcomes.

Additionally, since UTS and UAS addressed different content with varying levels of difficulty, the decrease in average scores should be seen as a descriptive observation rather than proof of academic decline. Future research is therefore recommended along several complementary directions. A quasi-experimental or randomized comparison design would enable stronger causal inferences about the effect of Edpuzzle on learning outcomes. Linking students' Edpuzzle engagement analytics- rewatch frequency and completion time- to subsequent achievement through multilevel modeling would help clarify how in-video behavior relates to summative performance. Baseline profiling of students' digital infrastructure and self-regulation capacity would allow environmental and dispositional factors to be tested rather than assumed, for example through validated self-regulated learning or learning-readiness scales, together with structured background surveys on participants' prior instructional exposure.

Finally, complementing quantitative measures with qualitative interviews or think-aloud protocols would illuminate the cognitive processes underlying performance shifts across the semester.

DECLARATION OF GENERATIVE AI USAGE IN THE WRITING PROCESS

During the preparation of this manuscript, the authors employed Claude as a generative AI writing assistant to support language refinement, proofreading, and the structural editing of the revised draft in response to the editor's and reviewers' comments. The AI tool was used solely for stylistic improvement, clarity of expression, and consistency of academic terminology; it was not used to generate original research data, statistical analyses, empirical findings, or interpretive conclusions, all of which remain the intellectual work of the authors. The authors have carefully reviewed, verified, and edited every portion of the AI-assisted content and take full responsibility for the accuracy, integrity, and originality of the published article.

REFERENCES

Alim, N., Linda, W., Gunawan, F., & Saad, M. S. M. (2019). The effectiveness of Google Classroom as an instructional medium: A case of State Islamic Institute of Kendari, Indonesia. *Humanities & Social Sciences Reviews, 7*(2), 240–246. https://doi.org/10.18510/hssr.2019.7227

Cesare, D. M. D., Kaczorowski, T., & Hashey, A. (2021). A piece of the (Ed)Puzzle: Using the Edpuzzle interactive video platform to facilitate explicit instruction. *Journal of Special Education Technology, 37*(3), 418–425. https://doi.org/10.1177/0162643421994266

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

Cornelison, B., Baldry, A., & Axon, D. R. (2022). Pharmacy students' experiences with an interactive video platform to develop clinical reasoning. *Pharmacy, 10*(4), 83. https://doi.org/10.3390/pharmacy10040083

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches (4th ed.)*. SAGE Publications.

Evans, J. D. (1996). *Straightforward statistics for the behavioral sciences*. Brooks/Cole Publishing.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

Field, A. (2018). *Discovering statistics using IBM SPSS Statistics (5th ed.)*. SAGE Publications.

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education (8th ed.)*. McGraw-Hill.

Galloway, K. R., Leung, M. W., & Flynn, A. B. (2018). A comparison of how undergraduates, graduate students, and professors organize organic chemistry reactions. *Journal of Chemical Education, 95*(3), 355–365. https://doi.org/10.1021/acs.jchemed.7b00743

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference (4th ed.)*. Allyn & Bacon.

Giannakos, M. N., Chorianopoulos, K., & Chrisochoides, N. (2016). Making sense of video analytics: Lessons learned from clickstream interactions, attitudes, and learning outcome in a video-assisted course. *The International Review of Research in Open and Distributed Learning, 17*(1), 1–23. https://doi.org/10.19173/irrodl.v16i1.1976

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis (8th ed.)*. Cengage Learning.

Hairida, H., Benõ, C., Soeharto, S., Charalambos, C., Rasmawan, R., Martono, M., Arifiyanti, F., Winarti, A., & Enawaty, E. (2023). Evaluating digital literacy of pre-service chemistry teachers: Multidimensional Rasch analysis. *Journal of Science Education and Technology, 32*(5), 643–654. https://doi.org/10.1007/s10956-023-10070-z

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice, 41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Mayer, R. E. (2009). *Multimedia learning (2nd ed.)*. Cambridge University Press.

Mayer, R. E. (2021). *Cognitive theory of multimedia learning*. In R. E. Mayer & L. Fiorella (Eds.), The Cambridge handbook of multimedia learning (3rd ed., pp. 57–72). Cambridge University Press.

Mischel, L. J. (2019). Watch and learn? Using Edpuzzle to enhance the use of online videos. *Management Teaching Review, 6*(3), 283–289. https://doi.org/10.1177/2379298118773418

Nsabayezu, E., Habimana, O., Nzabalirwa, W., & Niyonzima, F. N. (2025a). Optimizing pedagogical strategies in organic chemistry: Examining students' conceptual mastery and perspectives on the use of multimedia-supported flipped classroom approach. *Journal of Chemical Education, 102*(2), 456–467. https://doi.org/10.1021/acs.jchemed.4c01444

Nsabayezu, E., Habimana, O., Nzabalirwa, W., & Niyonzima, F. N. (2025b). Using multimedia-supported flipped classrooms approach to modernize organic chemistry instruction: Exploring students' and teachers' challenges. *Discover Education, 4*(1), 1–18. https://doi.org/10.1007/s44217-025-00738-6

Nsabayezu, E., Habimana, O., Nzabalirwa, W., & Niyonzima, F. N. (2025c). Contribution of activities-enhanced rubrics in the multimedia-supported flipped classroom approach on students' conceptual understanding in organic chemistry: A scoping review. *International Journal of Technologies in Learning, 32*(1), 44–61. https://doi.org/10.18848/2327-0144/CGP/v32i01/77-112

Okoth, R. (2025). Using Kahoot! as a formative assessment tool in undergraduate organic chemistry lecture courses. *Journal of Chemical Education, 102*(3), 780–789. https://doi.org/10.1021/acs.jchemed.4c01538

Pulukuri, S., & Abrams, B. (2020). Incorporating an online interactive video platform to optimize active learning and improve student accountability through educational videos. *Journal of Chemical Education, 97*(11), 3993–4002. https://doi.org/10.1021/acs.jchemed.0c00855

Rahiem, M. D. H. (2020). Technological barriers and challenges in the use of ICT during the COVID-19 emergency remote learning. *Universal Journal of Educational Research, 8*(11B), 6124–6133. https://doi.org/10.13189/ujer.2020.082248

Shelby, S. J., & Fralish, Z. D. (2021). Using Edpuzzle to improve student experience and performance in the biochemistry laboratory. *Biochemistry and Molecular Biology Education, 49*(6), 867–875. https://doi.org/10.1002/bmb.21494

Skulmowski, A., & Xu, K. M. (2022). Understanding cognitive load in digital and online learning: A new perspective on extraneous cognitive load. *Educational Psychology Review, 34*(1), 171–196. https://doi.org/10.1007/s10648-021-09624-7

Sukardi. (2011*). Metodologi penelitian pendidikan: Kompetensi dan praktiknya*. Bumi Aksara.

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review, 31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Van Gog, T., Hoogerheide, V., & van Harsel, M. (2020). The role of mental effort in fostering self-regulated learning with problem-solving tasks. *Educational Psychology Review, 32*(4), 1055–1072. https://doi.org/10.1007/s10648-020-09544-y

Weissgerber, T. L., Milic, N. M., Winham, S. J., & Garovic, V. D. (2015). Beyond bar and line graphs: Time for a new data presentation paradigm. *PLOS Biology, 13*(4), e1002128. https://doi.org/10.1371/journal.pbio.1002128

Weng, C., & Wirda, M. A. (2025). The effect of student's online learning readiness to student's engagement, satisfaction and achievement in higher education, Indonesia. *Education and Information Technologies, 30*(7), 9711–9736. https://doi.org/10.1007/s10639-024-13204-x

Zhang, D., Zhou, L., Briggs, R. O., & Nunamaker, J. F. (2006). Instructional video in e-learning: Assessing the impact of interactive video on learning effectiveness. *Information & Management, 43*(1), 15–27. https://doi.org/10.1016/j.im.2005.01.004