Comprehensive Journal of Science
Volume (11), Issue (42), (Sapt, 2026)
ISSN: 3104-6266

مجلة العلوم الشاملة
المجلد (11) العدد 42 سبتمبر 2026
ردمد: 3104-6266

Impact of Information Presentation on Undergraduates' Comprehension in Arabic Academic Information Environments

Emhimmid Elsanosi Algazairi [1*], and Dr. Anwar Alhenshiri [2]
1 Department of Electrical and Electronic Engineering, College of Technical Sciences, Misrata, Libya
2 Department of Computer Science, Faculty of Information Technology, Misrata University, Libya
*Corresponding author: Em.Elsanosi@gmail.com
Received: 12/10/2026 Revised: 15/10/2026 Accepted: 02/11/2026 Publication: 21/11/2026

أثر عرض المعلومات في فهم طلبة المرحلة الجامعية في بيئات المعلومات الأكاديمية العربية.

إمحمد السنوسي القزيري [1*]، د. أنور الهنشيري [2]
[1] قسم الهندسة الكهربائية والإلكترونية، كلية العلوم التقنية، مصراتة، ليبيا
[2] قسم علوم الحاسوب، كلية تقنية المعلومات، جامعة مصراتة، ليبيا

Abstract

This study compared three formats for presenting Arabic academic material and examined their relationship with immediate comprehension and students' experience of the task. Thirty-six undergraduates completed one task each on cloud computing, cybersecurity, and database systems. Each student used a different format for each topic: traditional prose, structured presentation, or structured presentation with learning aids. Participants were assigned in rotation to six sequences that varied the order of topics and formats. The analysis covered five outcomes across 108 tasks and used repeated-measures ANOVA with Greenhouse–Geisser correction and Holm adjustment. Mean comprehension scores were 2.42 for traditional prose, 3.83 for structured presentation, and 4.97 for the format with aids, out of six ($F = 42.157$, adjusted $p < .001$, partial $\eta^2 = .546$). With aids, students answered an average of 1.14 more questions correctly than with structured presentation alone (95% CI [0.62, 1.66]) and spent 37.22 seconds longer on reading and testing combined (95% CI [29.22, 45.23]). Compared with traditional prose, students answered 2.56 more questions correctly with aids and finished 9.31 seconds sooner. Effort, clarity, and user experience also differed across formats after adjustment. Thirty-five participants preferred the format with aids and one preferred structured presentation alone. These findings apply to the formats tested as a whole and to comprehension of the assigned material immediately after reading.

Keywords: structured information presentation; Arabic academic interfaces; comprehension; cognitive effort; user experience; repeated measures; learning aids.

الملخص:

بحثت الدراسة العلاقة بين طريقة عرض المادة الأكاديمية باللغة العربية وبين الفهم الفوري وتجربة الطلبة في أداء المهمة. شارك في الدراسة 36 طالبًا وطالبة من تخصصات الحوسبة. أنجز كل مشارك ثلاث مهام في الحوسبة السحابية والأمن السيبراني وقواعد البيانات، واستخدم في كل مهمة نمطًا مختلفًا من أنماط العرض الثلاثة: النص التقليدي، والعرض المنظّم، والعرض المنظّم المصحوب بمساعدات تعلّم. وُزّع المشاركون بالتناوب على ستة تسلسلات اختلف فيها ترتيب الموضوعات والواجهات. وشمل التحليل 108 سجلات للمهام وخمسة مؤشرات، باستخدام تحليل التباين للقياسات المتكررة مع تصحيح غرينهاوس–غايزر وضبط هولم.

بلغ متوسط درجات الفهم من ست درجات 2.42 للنص التقليدي، و3.83 للعرض المنظّم، و4.97 للعرض المصحوب بمساعدات التعلّم. وكانت الفروق دالة إحصائيًا بعد الضبط ($F = 42.157$، وقيمة p المعدّلة أقل من 0.001، ومربع إيتا الجزئي 0.546). وبالمقارنة مع العرض المنظّم وحده، أجاب الطلبة مع المساعدات عن 1.14 سؤال إضافي إجابة صحيحة في المتوسط (فاصل الثقة 95% من 0.62 إلى 1.66)، واستغرقوا 37.22 ثانية إضافية في القراءة والاختبار معًا (فاصل الثقة 95% من 29.22 إلى 45.23). وبالمقارنة مع النص التقليدي، ارتفع المتوسط بمقدار 2.56 إجابة صحيحة وانخفض متوسط زمن المهمة بمقدار 9.31 ثانية. كما اختلفت تقييمات الجهد المعرفي والوضوح وتجربة المستخدم بين الأنماط بعد الضبط. وفضّل 35 مشاركًا العرض المصحوب بالمساعدات، واختار مشارك واحد العرض المنظّم وحده. وتقتصر النتائج على الواجهات التي اختُبرت بمكوّناتها مجتمعة وعلى فهم المادة مباشرة بعد قراءتها.

الكلمات المفتاحية: العرض المنظم للمعلومات؛ الواجهات الأكاديمية العربية؛ الفهم؛ الجهد المعرفي؛ تجربة المستخدم؛ القياسات المتكررة؛ مساعدات التعلم.

1. Introduction

Students use academic information systems to locate documents and to understand the concepts they contain. They need to connect explanations, distinguish related ideas, and apply what they read to new situations. Search as Learning research examines learning outcomes alongside retrieval effectiveness (Urgo & Arguello, 2022, 2025). Finding a relevant document, however, does not establish how well a student understood it.

The organization of a text may help students understand it. Headings and grouped explanations show readers how ideas fit together. Prompts, glossaries, and self-check questions may encourage students to think further about those ideas, although these features also require additional reading and interaction. Research on hypertext and learning interfaces gives reasons to examine both possibilities (DeStefano & LeFevre, 2007; Faudzi et al., 2024). Whether students prefer a format and whether they understand more after using it are separate questions.

This study compared traditional prose, structured presentation, and structured presentation with learning aids using Arabic IT materials. The platform displayed the assigned material directly. Participants did not enter queries or retrieve live search results. Search as Learning informed the study, but the experiment examined how academic information was presented. Search effectiveness and the full search process were outside its scope.

The main question was whether the additional interaction involved in the format with learning aids was accompanied by better immediate comprehension. The study estimated differences in comprehension and time across the three Arabic formats and examined students' preferences and ratings separately. Each student used all three formats, reducing the influence of differences between students. Paired estimates are reported with their uncertainty. Arabic was the language of the study, and the design did not test whether the effect of presentation varies by language.

Section 2 reviews the relevant literature. Section 3 presents the research questions, objectives, and five hypotheses. Section 4 describes the implemented methodology, and Section 5 explains the analysis. Section 6 reports the results. Sections 7 and 8 discuss their contribution and limitations, and Section 9 concludes the paper.

2. Related Work

2.1 Search as Learning and Academic Information

In Search as Learning, the knowledge gained through an information system is one outcome of using it. Urgo and Arguello (2022) review approaches to assessing learning and discuss how measures can be selected and developed. Their review supports the use of a direct comprehension measure alongside ratings of satisfaction or confidence. Urgo and Arguello (2025) also trace how learning objectives, assessment, and search behavior have developed within the field.

The Search as Learning Spaceship model links learner characteristics, interfaces, retrieval systems, and information resources (von Hoyer et al., 2022). In the present experiment, all participants used the same platform and topic set. The presentation format varied. Since participants did not retrieve information through a live search, the findings cover the interface and the presentation of content within this model.

Zhao et al. (2025) examine patterns of search behavior associated with learning at different stages of a task. Niwanputri et al. (2025) examine prior knowledge and the cognitive demands of academic information searching. These studies informed the decision to record prior knowledge and the context of each interaction. They provide no basis for interpreting more clicks, scrolling, or mouse movement as evidence of better comprehension.

2.2 Information Presentation, Cognitive Effort, and User Experience

DeStefano and LeFevre (2007) review the cognitive demands of reading hypertext, including the effort required to navigate and integrate information. Faudzi et al. (2024) examine how to measure extraneous cognitive load in mobile learning interfaces. Both studies give reasons to examine the difficulty students report when using different formats. The validity of newly worded Arabic questionnaire items in the present setting still requires separate assessment.

Learning aids may draw attention to relationships in the material. Their value for learning, however, needs to be tested. In a controlled study of instructional video, Beege and Ploetzner (2025) found no improvement in learning performance from the prompts and navigation features they examined. Students reported greater cognitive demands on some measures when prompts were present. Although the medium differs, the study illustrates why both the benefits and the demands of learning aids deserve attention.

Research on signaling is more directly relevant to the organized displays. Schneider et al. (2018), in a meta-analysis of 103 studies, found benefits for retention and transfer when cues emphasized relevant information or its organization. These findings support using headings, highlighted terms, and conceptual links, although a package that includes them will not necessarily improve learning. The present aided condition also added explanations and self-check material, so it differs from a controlled manipulation of signaling alone.

Clarity, cognitive effort, and user experience capture different aspects of a task. Students may prefer a format without answering more questions correctly. They may also spend longer reading without finding the text more difficult. The study assessed these outcomes separately. Preference was recorded as part of the students' experience, while the test measured immediate comprehension.

2.3 Research Gap

Previous studies have examined visual cues, instructional prompts, and the difficulty learners report when using an interface. For students reading Arabic IT material, a practical question is whether comprehension scores and task times differ when learning aids are added to an already structured text. The present study examines these differences within the same group of students and considers their ratings of each format.

All three formats drew on the same source material. They differed in the information students could see and the aids they could use. The observed differences therefore apply to each format as a whole. They are useful for comparing the implemented interfaces, although the design cannot establish how much of a difference was due to headings, layout, or a particular aid.

3. Research Questions, Objectives, Hypotheses, and Significance

3.1 Research Questions

RQ1: How do students' immediate comprehension scores differ across the three formats used to present Arabic IT texts?

RQ2: How do traditional, structured, and structured-with-learning-aids presentation formats differ in terms of cognitive effort, perceived clarity, and user experience?

RQ3: How does the combined time for reading and completing the comprehension test differ across the three formats?

RQ4: What interaction patterns emerge across the three presentation formats?

3.2 Research Objectives

The main objective was to examine how presentation format relates to students' comprehension and learning experience when using Arabic academic material. The study compared comprehension scores, cognitive effort, clarity, user experience, and completion time. Recorded interactions provided additional descriptive evidence.

3.3 Research Hypotheses

H1: There is a statistically significant difference in comprehension scores across the three presentation formats.

H2: There is a statistically significant difference in perceived cognitive effort across the three presentation formats.

H3: There is a statistically significant difference in perceived clarity across the three presentation formats.

H4: There is a statistically significant difference in user experience across the three presentation formats.

H5: There is a statistically significant difference in task completion time across the three presentation formats.

The hypotheses were non-directional. Comprehension was the primary learning outcome. The remaining measures described other aspects of performance and task experience. The analysis was exploratory and was not preregistered. A correction for multiple testing was applied across the five omnibus tests.

3.4 Significance of the Study

The study compares three Arabic-language formats using a direct test of comprehension alongside students' preferences and ratings. For a designer considering a more elaborate interface, the comparison gives two relevant estimates: how much longer the task takes and how much the immediate test score differs. The findings apply to the three formats, the students who took part, and the test given immediately after reading.

4. Methodology

4.1 Research Design

The study used a within-subject repeated-measures design. Each participant completed three tasks, using each format once and studying each topic once. The topics were Cloud Computing, Cybersecurity, and Database Systems. Six fixed sequences varied the order of topics and formats. The platform assigned these sequences in rotation by participant number, without random allocation. Figure 1 shows the experimental framework.

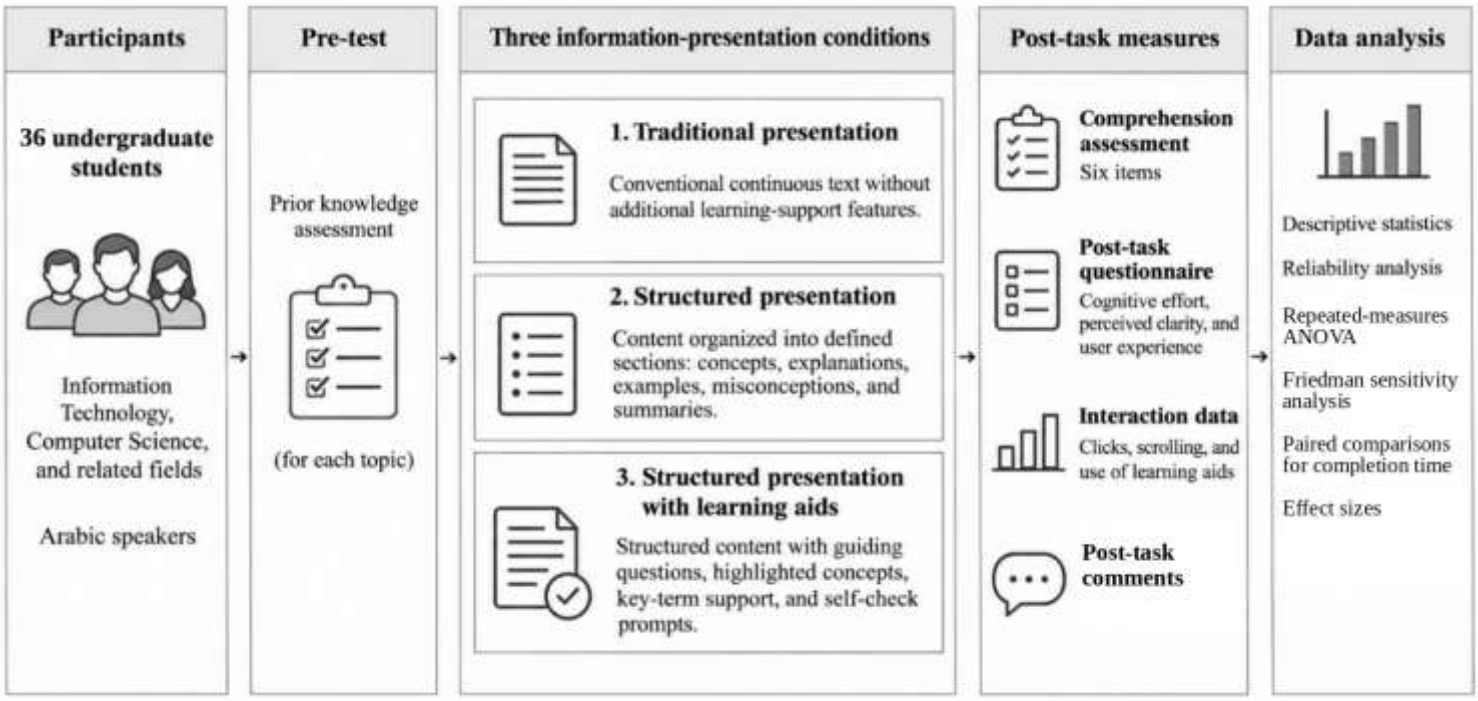


Figure 1. Experimental Framework and Evaluation Process

The independent variable was presentation format. The dependent measures were comprehension, task completion time, cognitive effort, clarity, and user experience. Topic, task position, and prior knowledge were recorded as contextual variables. Interaction measures were treated as supporting indicators rather than direct measures of attention or understanding.

4.2 Participants

The analyzed dataset contained 36 complete participant records and 108 tasks. Participants were undergraduates aged 18–25 years (M = 21.19, SD = 2.56). There were 22 women and 14 men. Majors were Information Systems (15), Software Engineering (8), Computer Science (7), and Information Technology (6). Students represented semesters three to eight. The convenience sample limits population-level inference.

All 36 participants included in the analysis completed tasks in all three formats. The exports did not show how many students were approached or identify all sessions that ended before submission. Participation and attrition rates could therefore not be calculated. The sample size also limits the precision of the estimates, particularly for small effects.

4.3 Experimental Conditions and Materials

The traditional condition presented a topic title followed by connected prose. The structured condition organized the material into labeled cards containing learning objectives, definitions, concepts, examples, and summaries. The aided condition combined structured cards with a guiding question, a reading path, highlighted terms, a concept diagram, and five expandable aid boxes. The boxes addressed terms, a key idea, conceptual links, points to remember, and an unscored self-check. Figure 2 illustrates the formats.

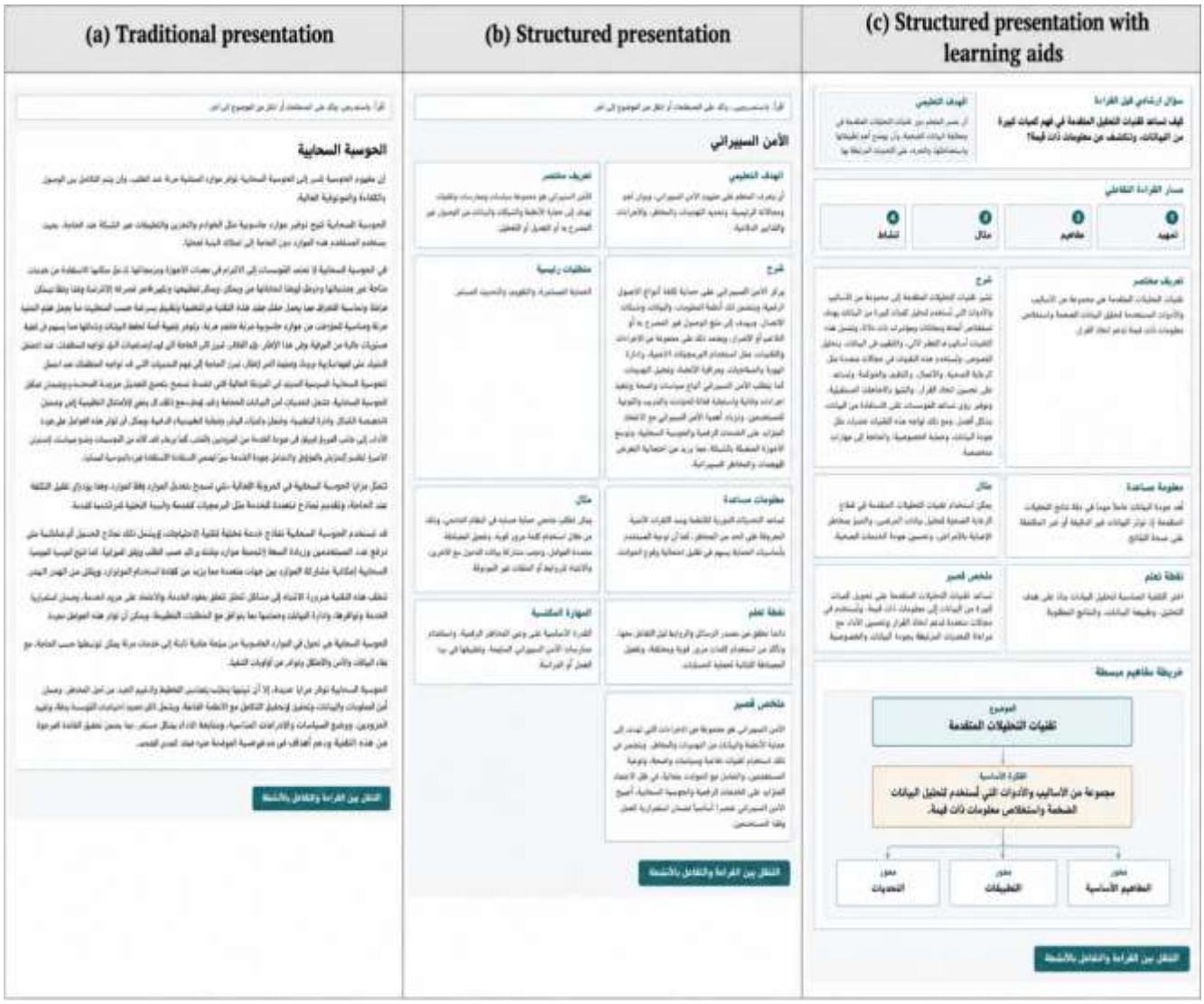


Figure 2. Examples of the Three Presentation Formats. The screens illustrate the layouts using different topics.

The three formats drew on the same source material, but students did not see exactly the same information in each one. For example, the structured cybersecurity screen in Figure 2 presents the definition and an example of account protection in separate cards. The traditional screen presents its topic in continuous prose. The format with aids included additional definitions and a self-check. The comparison therefore covers differences in both presentation and visible material. Each distinct aid box counted once, regardless of how often it was opened. The self-check was separate from the scored comprehension test.

4.4 Experimental Platform

The platform displayed Arabic material from right to left. It guided participants through consent, entry of participant information, instructions, tests, questionnaires, and the final choice of a preferred format. It also assigned participant codes and sequences, scored submitted answers on the server, and exported records by task and by participant. Students accessed the material directly without a search interface.

Participants had to answer all test questions and all required questionnaire items before moving to the next stage. They saw no scores during the tasks. After final submission, the platform stored the complete record and displayed the topic scores and total on the thank-you page. The detailed JSON records contained responses, timestamps, interaction events, and derived measures. The CSV files contained selected fields or summaries.

4.5 Measures

Table 1. Study Measures and Operational Definitions

Measure	Operational definition
Comprehension	Six scored responses per topic, summed from 0 to 6. Percentage = score / 6 × 100.
Prior knowledge	Three scored responses per topic, summed from 0 to 3.
Task completion time	Reading time plus comprehension-test time, in seconds. The pretest and questionnaire were excluded.
Cognitive effort	Mean of CE1–CE4, on a 1–5 agreement scale; higher values indicate greater reported effort or difficulty.
Perceived clarity	Mean of PC1–PC4, on a 1–5 agreement scale; higher values indicate clearer presentation.
User experience	Mean of UX1–UX5, on a 1–5 agreement scale; higher values indicate a more favorable experience.
Attention and preference	An attention item in task 2 required value 4. Final preferred task position was mapped to its actual interface.
Supporting interaction	Scrolling, clicks, answer revisions, and distinct aid-box use. Aid-open duration was not analyzed because it was stored as zero.

The prior-knowledge and comprehension tests used multiple-choice questions. Within each topic, the same six item identifiers and answer keys appeared across all three formats. The item records contained the selected option and indicated whether it matched the key. Each comprehension score was the number of correct answers on a six-question test. For each of the three questionnaire scales, the score was the mean of its items. Table 1 gives the interpretation of higher scores. Cronbach's alpha was used to assess internal consistency. The Arabic scales were not independently validated.

For each question, the recorded time ran from the opening of the test page to the last recorded selection for that question. It did not measure time spent on that item alone. The platform recorded reading time and test time separately. Their sum was the primary time outcome. Before analysis, the rounded durations were checked against the timestamps.

4.6 Counterbalancing and Procedure

Table 2 lists the six sequences. Participants were assigned to S1, S2, and the remaining sequences in order before the cycle began again. Each sequence included six participants. The six topic–format sequences covered three distinct orders of presentation format. Each format occurred 12 times in each task position and 12 times with each topic. This rotation balanced formats across topics and positions, although it did not randomize all six possible format orders.

Table 2. Implemented Sequences and Achieved Allocation

Sequence	Task 1	Task 2	Task 3	n
S1	Cloud B	Databases A	Cybersecurity C	6
S2	Cybersecurity B	Databases A	Cloud C	6
S3	Databases C	Cybersecurity B	Cloud A	6
S4	Cloud A	Cybersecurity C	Databases B	6
S5	Databases C	Cloud B	Cybersecurity A	6
S6	Cybersecurity A	Cloud C	Databases B	6

A = traditional prose; B = structured presentation; C = structured presentation with learning aids. Cloud = Cloud Computing.

Participants read the introduction, provided consent and demographic information, and reviewed brief instructions before the first task. Each task proceeded through the three-question pretest, reading, six-question comprehension test with no return to the material, and 13-item questionnaire. Task 2 included the attention check. After task 3, participants selected the first, second, or third interface they had experienced and gave a reason. The platform converted this position to the recorded format.

Recorded sessions began on 20 October 2026 between 11:00:12 and 14:46:14 Libya time (UTC+2). Session duration ranged from 1,349 to 1,838 seconds (22 min 29 s to 30 min 38 s), with a mean of 1,638.50 seconds (approximately 27 min 19 s).

4.7 Validity, Reliability, and Ethics

Participation was voluntary and informed consent was obtained. Analytical files used participant codes rather than names. All complete records were included in the analysis.

The checks covered participant identifiers, coverage of topics and formats, questionnaire ranges, consistency of scores, mapping of preferences to formats, and task times. Each reliability coefficient for an individual format was based on 36 participants. The pooled coefficients were based on 108 repeated task observations.

5. Data Analysis

The analysis treated each participant as the repeated-measures unit. Each of the 36 participants contributed one observation per format. The task export contained comprehension and pretest scores, questionnaire responses, and the duration of each phase. Task summaries were checked against the participant-wide export, and individual items were checked against the accompanying workbook.

Each of the five outcomes was compared using one-factor repeated-measures ANOVA. Greenhouse–Geisser epsilon was calculated from the centered within-participant covariance matrix, and the corrected degrees of freedom were used for all reported omnibus p values. Partial eta squared described the condition effect. Holm correction controlled the familywise error rate across the five omnibus tests (Holm, 1979). The significance threshold was .05. The analysis was finalized after data export and is reported as exploratory.

Because comprehension was a six-item bounded count and questionnaire scores were ordinal composites, Friedman analyses provided rank-based sensitivity evidence. Ranks were calculated with average ranks for ties and the tie correction (SciPy community, n.d.). Monte Carlo within-participant label permutations (99,999 per outcome; seed 20261104) supplemented the asymptotic reference distribution. Holm adjustment was applied across those five permutation p values. This is an exchangeability-based sensitivity analysis, not an allocation-based randomization test, because the actual assignment was cyclic.

Formal pairwise tests were limited to completion time. Holm correction was applied across its three paired t comparisons. Paired differences were reported with ordinary 95% confidence intervals and standardized paired effects (paired d). The confidence intervals were not adjusted for multiplicity. Paired comprehension differences and their ordinary 95% intervals were reported descriptively on the six-point scale.

An exploratory regression with participant fixed effects examined whether the differences in comprehension persisted after adjustment for topic, task position, and prior-knowledge score. The model used CR1 robust standard errors clustered by participant and t inference with 35 degrees of freedom. This analysis supplements the main comparison. Cyclic assignment and differences in the material displayed may still confound the estimates.

Calculations used Python 3.12.14, NumPy 2.3.5, pandas 2.2.3, and SciPy 1.17.0. The platform statistics export supplied descriptive summaries and repeated-measures F statistics. Confidence intervals, adjusted p values, and sensitivity analyses were calculated from the task observations.

6. Results

6.1 Data Completeness and Participant Characteristics

The dataset contained 36 distinct participant identifiers and 108 complete tasks. Together these records contained 324 pretest answers, 648 comprehension answers, and 1,404 responses to the core questionnaire items. Each task had a value for all five outcomes. All 36 participants passed the attention check in the second task.

Recomputed comprehension totals, percentage scores at the recorded precision, questionnaire means, and reading-plus-test durations agreed with the stored values. The core outcome and time fields also matched the wide

CSV. For the pretest, reading, test, and post-task survey phases, differences between start and end timestamps matched the recorded durations to the second. These checks confirmed that the exported files were internally consistent.

6.2 Descriptive Outcomes

Table 3. Outcome Means and Standard Deviations by Presentation Format

Outcome	Traditional	Structured	With aids
Comprehension (0–6)	2.42 (1.32)	3.83 (1.46)	4.97 (0.91)
Completion time (s)	338.61 (38.36)	292.08 (34.18)	329.31 (35.95)
Cognitive effort (1–5)	4.04 (0.51)	3.33 (0.52)	2.72 (0.54)
Clarity (1–5)	2.80 (0.68)	3.57 (0.50)	4.08 (0.46)
User experience (1–5)	2.83 (0.49)	3.51 (0.38)	4.06 (0.38)

Values are mean (sample SD); n = 36 per condition. Completion time includes reading and the comprehension test, but excludes the pretest and post-task questionnaire. Lower effort ratings indicate less reported difficulty.

Mean comprehension corresponded to 40.28% correct in traditional prose, 63.89% in structured presentation, and 82.87% in the aided format. Median scores were three, four, and five out of six, respectively. Mean prior-knowledge scores were 1.28, 1.19, and 1.28 out of three in the same order.

Mean clarity increased from 2.80 in traditional prose to 3.57 in structured presentation and 4.08 with aids. Mean reported effort decreased across the same formats (4.04, 3.33, and 2.72). The aided format also had the highest mean user-experience rating (4.06).

6.3 Omnibus Comparisons

Table 4. Repeated-Measures ANOVA with Corrected Inference

Outcome	F	GG epsilon	p GG	p Holm	Partial eta²
Comprehension (0–6)	42.157	0.899	< .001	< .001	0.546
Completion time (s)	78.114	0.953	< .001	< .001	0.691
Cognitive effort (1–5)	76.982	0.968	< .001	< .001	0.687
Clarity (1–5)	81.190	0.940	< .001	< .001	0.699
User experience (1–5)	139.852	0.905	< .001	< .001	0.800

Uncorrected degrees of freedom were 2 and 70 for every outcome. Greenhouse–Geisser corrected degrees of freedom equal these values multiplied by epsilon. p Holm adjusts the five p GG values; it does not represent an uncorrected platform p value.

The overall differences across formats remained significant for all five outcomes after Holm adjustment (all adjusted $p < .001$). For comprehension, $F = 42.157$, with GG-corrected degrees of freedom of 1.799 and 62.954 and partial $\eta^2 = .546$. Completion time also differed across formats ($F = 78.114$, partial $\eta^2 = .691$). These tests establish an overall difference. The paired comparisons below show which format had the higher score or longer time and the size of each difference.

Friedman sensitivity analyses gave the same overall pattern. For comprehension, $Q = 41.282$ and Kendall's $W = .573$. For completion time, $Q = 47.718$ and $W = .663$. All five Monte Carlo permutation p values were .00001 (.00005 after Holm adjustment). For cognitive effort, clarity, and user experience, the Friedman statistics were 55.674, 54.044, and 67.058, respectively, with corresponding Kendall's W values of .773, .751, and .931.

6.4 Completion Time and Comprehension Contrasts

Table 5. Paired Completion-Time Comparisons

Comparison	Difference (s)	95% CI (s)	p Holm	Paired d
Structured minus Traditional	-46.53	[-53.71, -39.34]	< .001	-2.19
Aids minus Traditional	-9.31	[-18.03, -0.58]	0.037	-0.36
Aids minus Structured	37.22	[29.22, 45.23]	< .001	1.57

Students completed the task 46.53 seconds sooner with structured presentation than with traditional prose. With learning aids, they finished 9.31 seconds sooner than with traditional prose but took 37.22 seconds longer than with structured presentation alone. All three time differences remained significant after Holm adjustment (Table 5). Mean reading times for traditional, structured, and aided presentation were 177.19, 127.19, and 164.19 seconds. The corresponding comprehension-test times were 161.42, 164.89, and 165.11 seconds.

The paired comprehension difference was 1.42 points for structured minus traditional (95% CI [0.76, 2.07]), 2.56 for aided minus traditional [2.04, 3.07], and 1.14 for aided minus structured [0.62, 1.66]. Figure 3 shows the score and time contrasts on their original scales.

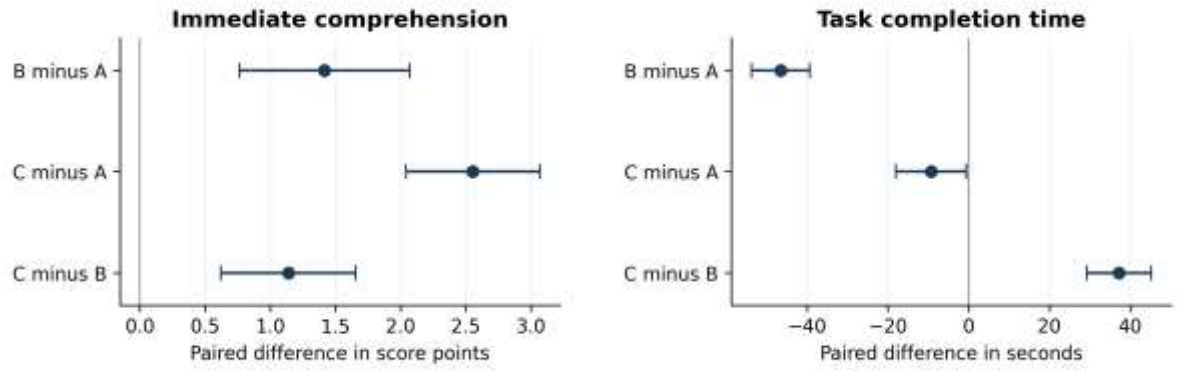


Figure 3. Paired Mean Differences with Ordinary 95% Confidence Intervals

A = traditional prose; B = structured presentation; C = structured presentation with learning aids. Intervals describe paired differences and are not multiplicity adjusted. The zero line denotes no mean difference.

6.5 Reliability, Preference, and Interaction

Table 6. Internal Consistency Estimated Separately by Condition

Scale	Traditional	Structured	With aids
Cognitive effort	0.759	0.700	0.672
Clarity	0.801	0.693	0.675
User experience	0.698	0.494	0.528

Across formats, alpha ranged from .672 to .759 for cognitive effort, from .675 to .801 for clarity, and from .494 to .698 for user experience (Table 6). Each estimate was based on 36 participants. Internal consistency for user experience was particularly low in the structured and aided formats, so ratings on this scale require caution. When the 108 repeated task records were pooled, the corresponding coefficients were .858, .863, and .833. Internal consistency alone does not establish translation validity or show that a scale measures a single dimension (Cronbach, 1951).

In the participant register, 35 participants (97.2%) selected the aided format and one (2.8%) selected the structured format as their final preference. None selected traditional prose. Across 108 tasks, the comment field contained 59 entries with nine distinct wordings. Only the counts are reported here.

In the format with aids, participants opened a mean of 3.39 distinct aid boxes (SD = 1.38, range 0–5). Mean scroll depth was 81.86% for traditional prose, 88.94% for structured presentation, and 82.61% for presentation with aids. Across the same formats, students changed their quiz answers an average of 0.89, 1.06, and 0.78 times per task. These measures describe recorded actions. The duration of aid use was excluded because the stored value of zero gave no measure of exposure.

Figure 4 shows the two components of task time and the final format preferences. Compared with structured presentation alone, students spent an additional 37.00 seconds reading with aids and 0.22 seconds more on the test. Compared with traditional prose, they spent 13.00 fewer seconds reading and 3.69 seconds more on the test. These figures describe how the mean time was divided. Each participant recorded one final preference.

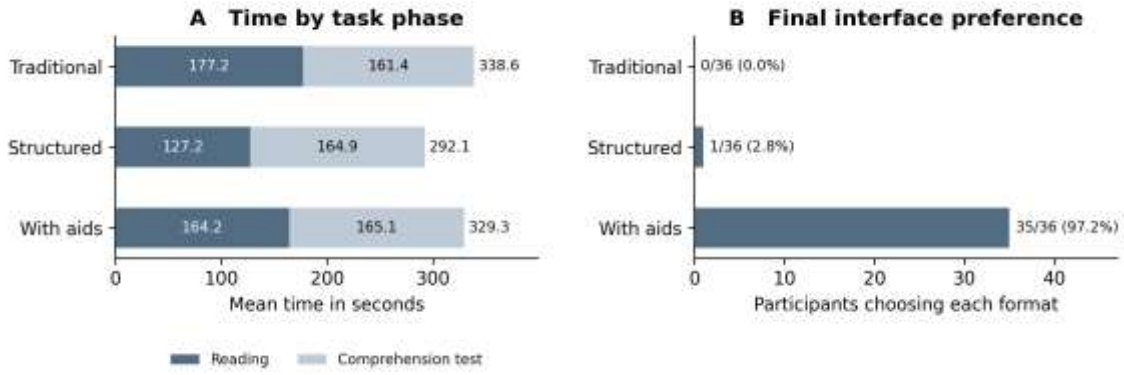


Figure 4. Task Time Components and Final Interface Preference

Panel A shows mean reading and comprehension-test times for 36 observations per format; the totals exclude pretests and questionnaires. Panel B shows one final preference per participant (n = 36). Both panels are descriptive and do not imply a participant-level relationship between time and preference.

6.6 Sensitivity Analyses

After adjustment for topic, task position, and prior knowledge, the estimated comprehension difference was 1.45 points for structured minus traditional (cluster-robust 95% CI [0.63, 2.26]), 2.56 for aided minus traditional [1.94, 3.17], and 1.11 for aided minus structured [0.49, 1.73]. The direction of all three estimates agreed with the unadjusted paired differences.

7. Discussion and Research Contributions

7.1 Immediate Comprehension and Time across Presentation Formats

Mean immediate comprehension was highest for the format with learning aids. The time difference depended on the comparison. Students spent about 37 seconds longer on the aided task than on the structured task, but finished about nine seconds sooner than with traditional prose. Structured presentation alone had the shortest combined reading and test time.

Almost all the additional time with aids, compared with structured presentation alone, was spent in the reading phase. The aided format included extra explanations and a self-check. The data cannot distinguish time spent on these features from differences due to layout or navigation. Beege and Ploetzner (2025) reported a different pattern for learning performance in instructional video. Their video experiment controlled learning time, whereas the present study recorded how long students spent in each format. This difference, along with the different media and aids, limits direct comparison.

7.2 Preference and Comprehension

Of the 36 participants, 35 preferred the format with learning aids. This format also had the highest mean comprehension score. Agreement at the group level does not establish whether an individual student's preference predicted learning. A test score and a preference rating describe different aspects of the interface (Urgo & Arguello, 2022).

On average, participants reported less effort, greater clarity, and a better user experience with learning aids. The overall differences across formats remained significant after adjustment. Of the three scales, user experience had the weakest internal consistency within formats. This limits the precision with which its combined rating can be interpreted.

7.3 Scope of the Empirical Contribution

Assessing comprehension, time, and preference in the same students gives a fuller account of how the formats were used. The format most students preferred also had the highest mean test score, while structured presentation alone had the shortest mean completion time. The rank-based analysis supported the overall differences, and estimating reliability within each format showed variation that was less apparent in the pooled results. These findings apply to the materials tested and to comprehension immediately after reading.

7.4 Implications for Interface Design and Replication

For a designer choosing between the two structured formats, the version with aids had a higher immediate test score and required about 37 seconds more per task. Where reading time is limited, structured presentation alone was the fastest of the three formats. Testing the aids separately would help establish which ones to include.

A confirmatory study could present the same testable points in each format, vary individual aids, and randomize assignment across a fully balanced set of sequences. Specifying the smallest educationally meaningful score difference in advance would support sample-size planning and equivalence testing. A delayed test would show whether students retained the material beyond the immediate assessment, a question left open by the present study.

8. Limitations and Future Work

The convenience sample included 36 participants from four computing-related majors. With only six questions per topic, the immediate comprehension test could distinguish only a limited range of scores. The study did not establish content validity, the validity of the Arabic adaptations, or whether each scale measured a single dimension. The relatively low internal consistency of the user-experience scale within formats adds uncertainty to its interpretation.

The pretest and the comprehension test differed in length and were not equated. Subtracting their raw scores would therefore not give a measure of learning gain. The six-question tests still require evidence that they cover the intended content, that item difficulty is understood, and that the demands are comparable across topics. Adjusting for topic accounts for measured mean differences but cannot validate the tests.

Each topic–format combination and each format–position combination occurred 12 times. The six cyclic sequences nevertheless covered only three distinct format orders. Since participants were assigned without randomization, unmeasured sequence or carryover effects may remain. The adjusted analysis accounts for measured differences in topic, task position, and prior knowledge.

Students saw different information across the three formats. The effects of layout and instructional content therefore cannot be fully separated. The platform displayed the assigned texts directly and tested comprehension immediately. These results have limited application to unrestricted search, retention, or broader academic performance. Interaction events were sampled, the mouse measures covered different parts of the task, and the stored aid-open duration of zero could not measure exposure. The interaction results are therefore descriptive.

The final analysis was conducted after data export and was exploratory. Adjustment for multiple tests and checks using alternative analyses limited some analytical choices. An independent study with a protocol registered in advance would provide a stronger test of the observed pattern.

Future work should plan for a larger sample, use random allocation, balance all format orders, and match the material across formats on the points covered by the test. A delayed test could show whether the immediate differences in comprehension persist. Testing glossaries, prompts, diagrams, and self-checks separately could clarify the contribution of each aid. Stronger conclusions about the mechanisms of cognitive load would also require validated instruments.

9. Conclusion

The 36 undergraduates completed 108 tasks. Their mean immediate comprehension score was highest with structured Arabic presentation and learning aids, and 35 participants preferred this format. Compared with structured presentation alone, students answered an average of 1.14 more questions correctly out of six with aids

and spent 37.22 seconds longer on reading and testing. Compared with traditional prose, they answered 2.56 more questions correctly and finished 9.31 seconds sooner. These results apply to the formats as a whole. A randomized replication with matched materials and a delayed test is needed to identify which components account for the differences and whether the advantage persists.

References

Beege, M., & Ploetzner, R. (2025). Learning from interactive video: The influence of self-explanations, navigation, and cognitive load. Instructional Science, 53, 99–119. https://doi.org/10.1007/s11251-024-09693-5

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297–334. https://doi.org/10.1007/BF02310555

DeStefano, D., & LeFevre, J.-A. (2007). Cognitive load in hypertext reading: A review. Computers in Human Behavior, 23(3), 1616–1641. https://doi.org/10.1016/j.chb.2005.08.012

Faudzi, M. A., Che Cob, Z., Ghazali, M., Omar, R., & Sharudin, S. A. (2024). User interface design in mobile learning applications: Developing and evaluating a questionnaire for measuring learners' extraneous cognitive load. Heliyon, 10(18), e37494. https://doi.org/10.1016/j.heliyon.2024.e37494

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6(2), 65–70. https://doi.org/10.2307/4615733

Niwanputri, G. S., Tob-Ogu, A., & Hosseinzadeh, M. (2025). Untangling cognitive processes in academic information searching: Investigating tool gaps, cognitive load, and user satisfaction. In Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR '25) (pp. 449–458). Association for Computing Machinery. https://doi.org/10.1145/3731120.3744615

Schneider, S., Beege, M., Nebel, S., & Rey, G. D. (2018). A meta-analysis of how signaling affects learning with media. Educational Research Review, 23, 1–24. https://doi.org/10.1016/j.edurev.2017.11.001

SciPy community. (n.d.). scipy.stats.friedmanchisquare. SciPy documentation. https://docs.scipy.org/doc/scipy-1.17.0/reference/generated/scipy.stats.friedmanchisquare.html

Urgo, K., & Arguello, J. (2022). Learning assessments in search-as-learning: A survey of prior work and opportunities for future research. Information Processing & Management, 59(2), 102821. https://doi.org/10.1016/j.ipm.2021.102821

Urgo, K., & Arguello, J. (2025). Search as learning. Foundations and Trends in Information Retrieval, 19(4), 365–556. https://doi.org/10.1561/1500000084

von Hoyer, J., Hoppe, A., Kammerer, Y., Otto, C., Pardi, G., Rokicki, M., Yu, R., Dietze, S., Ewerth, R., & Holtz, P. (2022). The Search as Learning Spaceship: Toward a comprehensive model of psychological and technological facets of search as learning. Frontiers in Psychology, 13, 827748. https://doi.org/10.3389/fpsyg.2022.827748

Zhao, Y., Huang, H., Li, Q., & Zhao, Y. (2025). How user information search patterns impact learning outcomes across different stages of learning-related search tasks. Library & Information Science Research, 47(2), 101354. https://doi.org/10.1016/j.lisr.2025.101354