International Journal of Evaluation and Research in Education (IJERE)
Vol. 15, No. 4, August 2026, pp. 3172~3181
ISSN: 2252-8822, DOI: 10.11591/ijere.v15i4.39915

ChatGPT as scaffold: quiz performance across session complexity

Fatima Ezzahra Kabba[1,2], Zouhair Ejbari[2]
[1]Higher International Institute of Tourism (ISITT), Ministry of Tourism, Handicrafts, and Social and Solidarity Economy, Tangier, Morocco
[2]Department of Economics and Management, Faculty of Legal and Economic Sciences of Tangier, Abdelmalek Essaadi University, Tangier, Morocco

Article Info

Article history:

Received Apr 12, 2026
Revised Jul 11, 2026
Accepted Jul 20, 2026

Keywords:

ChatGPT
Generative AI
Higher education
Quasi-experimental design
Statistics education
Student performance

ABSTRACT

Many studies examine the use of ChatGPT in education, but most measure student perceptions, not performance, and few track performance across multiple sessions of different complexity. In particular, no study has tracked whether this association varies across sessions of different cognitive complexity. Using a quasi-experimental design, first-year undergraduates (N=193) at the Higher International Institute of Tourism (ISITT) in Tangier, Morocco were followed across seven introductory statistics sessions. One group had access to ChatGPT during learning activities, while the other followed the same instruction without artificial intelligence (AI) access. Performance was measured through end-of-session quizzes (1,136 observations) and analyzed using a linear mixed-effects model. No consistent overall advantage was associated with either condition. However, a significant interaction between condition and session was identified ($\chi^2(6)=61.50$, $p<0.001$). The most pronounced divergence occurred in session 6, which involved multi-step computation (harmonic mean and grouped-data mode interpolation), where the AI-permitted group scored higher (9.15 vs. 7.18, $d=1.24$). Across the remaining six sessions, effect sizes ranged from $d=-0.20$ to $d=+0.19$. Our findings suggest that AI integration should be selective, used during complex procedural sessions rather than uniformly across all course content.

Corresponding Author:

Fatima Ezzahra Kabba
Higher International Institute of Tourism (ISITT)
Ministry of Tourism, Handicrafts, and Social and Solidarity Economy
Tangier, Morocco
Email: kabba.fatima.ezzahra@gmail.com

1. INTRODUCTION

Generative artificial intelligence (AI) tools such as ChatGPT are becoming increasingly common in higher education, as they can help students in many ways: generating explanations, examples, and step-by-step guidance in response to problems they face in learning [1]–[4]. Despite this growing use, teachers still lack clear guidance on when and how to use these tools in their courses. A practical question remains largely unresolved: under what conditions, if any, can these tools support student learning when used alongside regular classroom instruction?

Most existing studies focus on student perceptions, ethical debates, or institutional policies, with fewer examining actual learning outcomes across multiple sessions [2], [5]–[7]. The studies that do measure performance tend to rely on single sessions, small samples, or self-reported data [8]. Longitudinal classroom-based evidence, tracking real student performance as course content grows in complexity, remains limited. This gap matters in quantitative disciplines such as statistics, where learning builds cumulatively and procedural demands increase from one session to the next. To our knowledge, no prior study has examined whether ChatGPT support is more strongly associated with student performance in procedurally complex sessions than in conceptual ones, or how this association varies across sessions of differing cognitive complexity.

This gap is especially relevant in North African countries like Morocco, where digital access is unequal and institutional AI policies do not yet exist. Students and teachers navigate AI integration without pedagogical guidance or institutional frameworks [9], [10]. A recent review confirms that while ChatGPT has gained wide recognition among students, its limitations require careful attention, especially in settings with limited institutional support [11]. The specificity of this study also lies in its institutional and pedagogical context. Many students at Higher International Institute of Tourism (ISITT) begin the descriptive statistics course with little or no prior statistical background, particularly those who come from a humanities baccalaureate track. Descriptive statistics is a mandatory first-semester course, making it a cognitively demanding component of their academic transition. Having taught this subject for more than a decade, the first author observed that students often require substantial support to move from intuitive reasoning to statistical thinking. Furthermore, the study was conducted in the absence of a formal institutional AI policy, a situation that reflects the broader uncertainty surrounding the governance of generative AI use in Moroccan public higher education. Research from Middle East and North Africa (MENA) teachers confirms that AI adoption in the region faces unique challenges related to institutional readiness and training [12]. A global survey of over 23,000 students from 109 countries found that most ChatGPT research comes from Western contexts [13].

While these studies describe who uses ChatGPT and how it is perceived, they do not explain when AI helps and when it does not. A framework from mathematics education helps address this question. Hiebert and Lefevre [14] differentiate between conceptual knowledge (understanding what ideas mean) and procedural knowledge (the ability to carry out step-by-step actions). For example, understanding why the median is preferred for skewed distributions is conceptual. Applying the interpolation formula to find the mode of grouped data is procedural. Star [15] adds that procedural knowledge can range from rote execution to flexible application. In this study, conceptual tasks are those where students define, classify, or interpret statistical concepts. Procedural tasks are those where students select and apply formulas through multi-step computation.

In education, scaffolding refers to temporary instructional support that helps students complete tasks they cannot yet do independently, and that is reduced step by step as the student becomes more capable. When used to verify intermediate steps or break down complex problems, ChatGPT can function as a form of scaffolding that supports learners at moments of difficulty [16]. This kind of support may be most relevant during procedural transitions, where students must move from understanding concepts to applying formulas, consistent with active learning approaches that emphasize verification and reformulation [17]. Recent research indicates that generative AI may act as a 'more knowledgeable other' in Vygotsky's framework, scaffolding learning within the zone of proximal development [18]. Yet, continuous AI availability may also lead to over-scaffolding, limiting students' development of independent problem-solving skills [19]. This raises a key question: does AI support learning primarily when tasks are beyond what students can do without help, or does it create dependency even in less demanding tasks? This study tests whether ChatGPT access matters differently across sessions of varying complexity. The idea is simple: ChatGPT access may help more in some sessions than others, depending on how difficult the task is. Figure 1 shows this conceptual model.

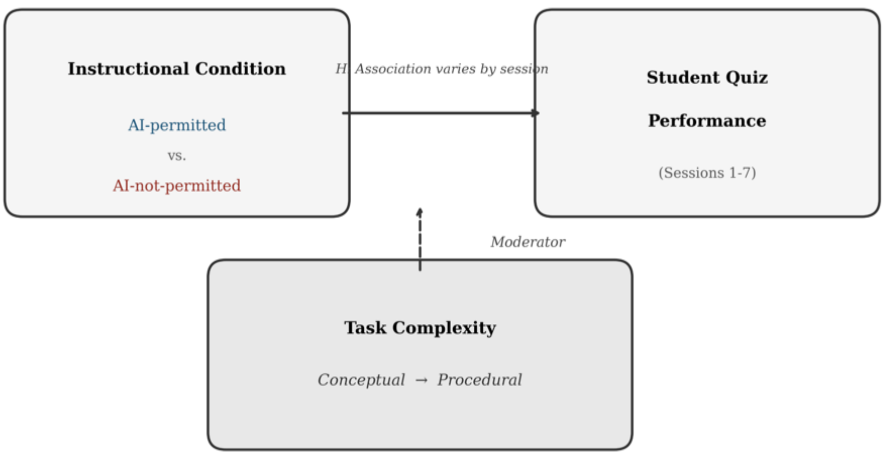

Figure 1. Association between instructional condition and performance, moderated by task complexity

The study examines student performance across seven sessions in which one group used ChatGPT alongside teacher-led instruction during learning activities, while a second group followed the same instruction without AI access. To the authors' knowledge, no classroom-based quasi-experimental study on ChatGPT and student performance has been conducted in a Moroccan higher education institution. This paper provides classroom-based longitudinal evidence from a context where such empirical research remains scarce.

The novelty of this study lies in three aspects. First, it is, to our knowledge, the first quasi-experimental longitudinal study tracking the association between ChatGPT access and student quiz performance across sessions of different cognitive complexity. Second, unlike studies that measure perceptions or rely on single-session designs, this study provides empirical evidence that the association is session-specific, emerging specifically at a procedural complexity threshold. Third, it contributes classroom-based evidence from a Moroccan higher education institution, where such empirical research remains scarce.

2. METHOD

This section describes the research design, participants, teaching approach, and data analysis procedures used in the study. The study adopted a quasi-experimental classroom-based design to examine student quiz performance across seven instructional sessions. All procedures were conducted under normal teaching conditions at ISITT in Tangier, Morocco.

2.1. Study design

This study adopted a quasi-experimental, classroom-based design conducted under normal teaching conditions. Instructional conditions were determined by pre-existing class organization rather than random assignment. The study was conducted within regular course delivery and did not involve changes to the institutional timetable or external assessments. The study examines associations between instructional condition and performance rather than making causal claims.

The study was conducted over seven consecutive sessions beginning on November 21st, 2024, each addressing a distinct set of introductory statistical concepts and concluding with a short in-class quiz. Each session focused on a different set of introductory statistical concepts, aligned with the course syllabus. Sessions followed a consistent structure: introduction of concepts, guided practice, and an end-of-session quiz.

2.2. Participants

Participants were first-year students enrolled in an introductory statistics course at the ISITT in Tangier, Morocco. Students were organized into five pre-existing classes. One class was divided into two separate groups, producing six groups, three following AI-permitted instruction and three following AI-not-permitted instruction. Table 1 summarizes the arrangement.

The conditions represent permission-based learning environments, not verified AI tool adoption. The study did not measure or verify whether individual students actually used ChatGPT. Analyses are interpreted as associations between condition and performance. Across seven quiz sessions, 193 unique students submitted at least one quiz, yielding 1,136 observations.

Table 1. Study design: class and group arrangement

Class	AI-permitted (3 groups)	AI-not-permitted (3 groups)
Class 1	Full class assigned	—
Class 2	Full class assigned	—
Class 3	—	Full class assigned
Class 4	—	Full class assigned
Class 5 (split)	Group a–separated	Group b–separated
Total observations	563	573

N=193 unique students, 1,136 quiz observations across seven sessions.

2.3. Teaching approach and use of ChatGPT

All classes followed the same planned content, slides, and exercises. In AI-permitted classes, students could consult ChatGPT during learning activities only. Students were required to rewrite all explanations in their own words. The instructor monitored interactions and corrected AI-generated inaccuracies when observed. In AI-not-permitted classes, students relied on lecture materials and independent reasoning. ChatGPT was not allowed during quizzes in any class.

2.4. Quizzes

At the end of each session, students completed a quiz aligned with that session's learning objectives. Scores were recorded on a 0–10 scale. Quizzes differed in content and cognitive demands across sessions. The model includes a separate term for each session, which controls for the fact that some quizzes were harder than others.

Three examples illustrate the range. In quiz 2 (conceptual), students distinguished between population, sample, statistical unit, and variable, a task that proved challenging because many students interpreted 'population' only in the everyday sense. In quiz 4 (procedural), students computed absolute and relative frequencies from a raw dataset. In quiz 6 (complex procedural), students calculated the harmonic mean from a frequency table using H=N/Σ(n_i/x_i), and determined the mode of a grouped continuous variable by computing density for each class, identifying the modal class, and applying the interpolation formula. The full quiz 6 items are reported in Appendix.

2.5. Data analysis

A linear mixed-effects model was used with fixed effects for instructional condition, session (categorical), and their interaction, and a random intercept for student ID. The model is specified in (1). A linear mixed-effects model was chosen because students contributed between one and seven observations depending on attendance, producing an unbalanced repeated-measures structure [20]. This model accounts for individual differences through a random intercept and does not require equal numbers of observations per student. This approach was preferred over repeated-measures analysis of variance (ANOVA) because ANOVA requires complete data for all sessions from each student, which was not available due to natural attendance variation. The mixed-effects model uses all available observations without excluding students who missed one or more sessions.

$$Score_{ij} = \beta 0 + \beta 1(Condition) + \beta 2(Session) + \beta 3(Condition \times Session) + u\_i + \varepsilon\_ij \quad (1)$$

Here, u_i denotes the random intercept for student i, and ε_ij is the residual error for observation j within student i.

We used Python (Anaconda) with the statsmodels library and restricted maximum likelihood (REML) estimation. The intraclass correlation coefficient (ICC) from the unconditional model was 0.034, indicating that approximately 3% of the total variance in quiz scores was attributable to stable differences between students. The remaining 97% reflected within-student variation across sessions. The multilevel structure is warranted to account for repeated observations within students.

3. RESULTS AND DISCUSSION

This section presents the descriptive statistics, model estimates, and session-level findings from the linear mixed-effects analysis. Results are organized from overall group-level patterns to specific session effects, with a focus on the interaction between instructional condition and session complexity. The discussion integrates these findings with existing literature and addresses their practical and theoretical implications.

3.1. Descriptive overview

Table 2 presents the descriptive statistics for both groups, including sample size, mean quiz score, and standard deviation (SD). The two groups showed similar overall performance when scores were combined across all seven sessions. These aggregate figures provide a baseline before examining session-level variation in subsequent subsections.

The two groups had similar mean scores (7.30 vs. 7.06) and similar variability (SD=2.39 vs. 2.29) when performance was combined across all sessions. These aggregate averages may mask important session-level variation, which is examined in the following sections. The comparable group-level statistics also indicate that any observed differences are unlikely to result from a general performance gap between the two groups.

Table 2. Descriptive statistics by instructional group (N=1,136)

Group	n	Mean	SD
ChatGPT-permitted	563	7.30	2.39
No ChatGPT	573	7.06	2.29

3.2. Overall effects

A joint Wald test (which evaluates whether a set of model coefficients is significantly different from zero) showed no significant overall effect of instructional group ($\chi^2(1)=1.07$, $p=0.301$). However, the interaction between group and session was statistically significant ($\chi^2(6)=61.50$, $p<0.001$), indicating that the association was not the same in every session. Table 3 reports the fixed-effects estimates.

Table 3. Fixed-effects estimates from the linear mixed-effects model

Effect	Estimate	SE	z	p
Intercept	4.71	0.19	25.06	<0.001
Group (ChatGPT)	0.27	0.26	1.04	0.301
Session (overall)	—	—	—	<0.001
Group×Session	—	—	—	<0.001

3.3. Session-level results

Table 4 presents the mean quiz scores, sample sizes, and Cohen's d effect sizes for each group across all seven sessions. These results allow direct comparison of both aggregate and session-specific performance patterns between conditions. Figure 2 visualizes the score trajectories for both groups across sessions.

Table 4. Mean quiz scores, sample sizes, and effect sizes by session

S	Topic	n(AI)	M(AI)	n(No)	M(No)	d	Type
1	Descriptive vs. inferential	81	4.99	80	4.75	+0.18	Conceptual
2	Statistical vocabulary	78	7.62	81	7.68	−0.03	Conceptual
3	Variable types	83	8.31	86	8.42	−0.05	Conceptual
4	Frequencies	80	9.65	85	9.48	+0.19	Procedural
5	Arithmetic/geometric means	81	5.23	83	5.59	−0.20	Procedural
6	Harmonic mean and mode	78	9.15	83	7.18	+1.24	Complex procedural
7	Graphs and median	82	6.27	75	6.05	+0.12	Application

d=Cohen's d (positive=AI group higher). Classification follows [14].

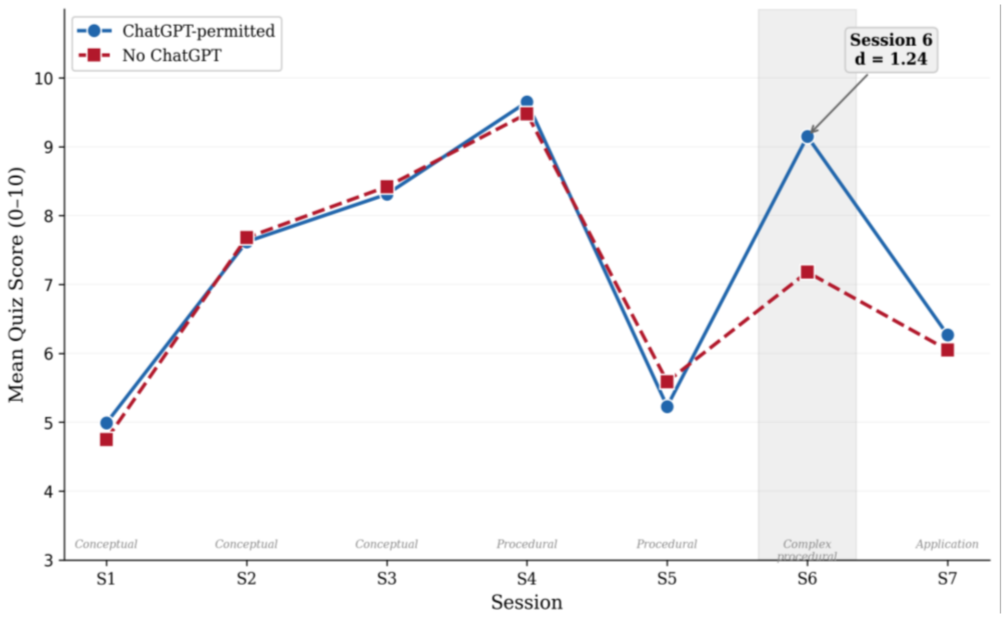

Figure 2. Mean quiz scores across seven sessions for both groups. The two lines overlap closely in sessions 1–5 and 7, diverging only in session 6

3.4. Session 6 divergence

The largest difference was in session 6: the AI-permitted group scored 9.15 compared with 7.18 ($\beta=1.67$, standard error (SE)=0.34, $z=4.89$, $p<0.001$). The standardized effect size was $d=1.24$, calculated as the raw mean difference (1.97) divided by the pooled within-session SD (1.59). Following Cohen [21], this represents a large effect. This effect emerged in only one of seven sessions; across the remaining six, effect sizes ranged from $d=-0.20$ to $d=+0.19$. Figure 3 shows the effect sizes for all sessions.

Session 6 represents the highest level of procedural demand in the course. Students had to select the correct formula, compute density for each class, identify the modal class, and apply an interpolation procedure. The higher scores appeared at this moment of accumulated procedural complexity, not uniformly across the course.

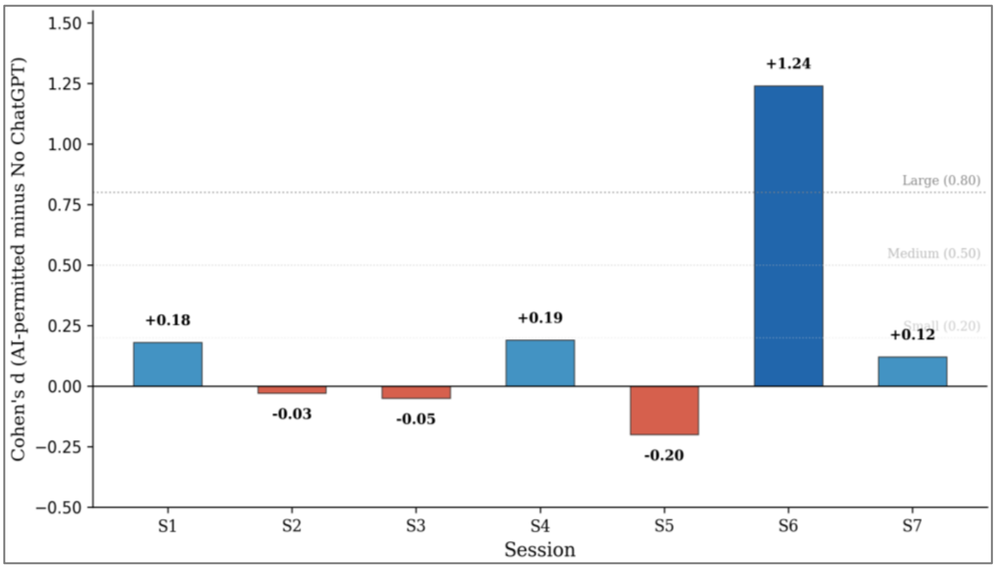

Figure 3. Cohen's d effect sizes by session. Dashed line=large-effect threshold (0.80)

3.5. Discussion

Instructional conditions were permission-based and not experimentally controlled. All findings are therefore interpreted as conditional associations rather than causal effects. The following discussion considers what the observed patterns suggest, while acknowledging the design limitations inherent to this classroom-based study.

The first five sessions focused on foundational concepts and both groups performed similarly. In session 6, students faced multi-step computation, identifying a modal class based on density, applying an interpolation formula, and sequencing several computational steps. Under these conditions, the AI-permitted group scored notably higher (d=1.24). This finding is consistent with Alneyadi and Wardat [22], who reported positive effects of ChatGPT on student achievement in a procedurally demanding science unit. A recent quasi-experimental study in a tourism statistics course similarly found that ChatGPT's association with performance depends on the specific task context [23].

The score drop-in session 5 provides important context. Both groups experienced a score drop when the course shifted from definitional content to formula-driven computation. Yet the groups did not diverge in session 5, the divergence came only in session 6. In other words, ChatGPT may not help when students first encounter formulas, but it may start to matter when the formulas become too complex to handle without support. In contrast, Yang *et al*. [24] found that unsupervised ChatGPT use decreased self-efficacy and achievement in programming. The difference may be explained by the level of instructor oversight: in their study, students used AI independently, whereas in the present study, use was monitored and students were required to reformulate AI responses in their own words.

Xing [25] showed that ChatGPT performs well on conceptual questions. Other studies have reported positive effects in specific subject areas, including remote learning in medical education [26] and knowledge gains in music education [27]. Our study goes further by showing that the association between AI access and student performance appears specifically during procedural sessions, not during conceptual ones. The concerns about AI bypassing standard procedures [28] remain relevant, in this study, instructor oversight may have limited this risk. The broader literature suggests that structured AI use is more likely to support learning than unrestricted access [29], [30]. The session-specific pattern may also be understood through cognitive load theory, which describes how the mental effort required by a task affects learning [31]. Sessions 1–3 required students to remember definitions and classify concepts, tasks that do not demand much mental effort. Session 6 required multi-step computation with several formulas applied in sequence, a task with high mental effort, where ChatGPT may have helped by breaking the problem into smaller steps [32]. However, constant AI availability also raises concerns. Students may stop trying to solve problems on their own if AI always gives them the answer [19]. In this study, ChatGPT access did not improve scores in sessions 1–3, which means students did not become dependent on AI for tasks they could handle alone. A recent meta-analysis confirms that ChatGPT's impact on academic achievement is not uniform but varies by course type and duration [33]. These results challenge the assumption that AI produces universal benefits in education. The present study adds that even within a single course, the association varies by session.

For statistics instructors, the approach used in this study offers a concrete model: explain the procedure without AI, allow ChatGPT during practice for step verification, require reformulation in the student's own words, and prohibit AI during the quiz. This sequence was associated with stronger performance at the point of highest procedural demand. It should be treated as a preliminary step, not a verified method.

When all seven sessions were considered together, no consistent advantage was associated with AI availability. The effect was limited to session 6. For sessions focused on definitions and classifications, teacher-led instruction without AI appeared equally effective. AI integration should be selective, not uniform.

3.6. Limitations and future research

Several design features limit the conclusions. First, the groups reflected permission-based conditions rather than verified AI use—the study did not collect prompt logs or screen recordings to confirm whether students actually used ChatGPT or how often. Second, the split-class arrangement in one class means some informal cross-condition interaction cannot be fully excluded, though any contamination would be expected to reduce rather than inflate observed differences. Third, the freely available version of ChatGPT occasionally produced inaccurate explanations that required instructor correction, which reflects real-world constraints but introduces variation that was not systematically recorded. Fourth, the study was conducted at a single institution in Morocco with a specific curriculum, a single instructor, and first-year students, limiting generalizability to other institutions, disciplines, or student populations. Future research should include direct measures of AI interaction such as prompt logs collected with student consent, and should use external experts to rate session complexity independently.

4. CONCLUSION

This study examined the association between ChatGPT access and student quiz performance across seven sessions in an introductory statistics course at the ISITT in Tangier, Morocco. A significant interaction between group and session was identified, indicating that the association was session-specific rather than uniform. It contributes classroom-based longitudinal evidence from a Moroccan higher education institution, a context where such empirical research remains scarce.

The clearest pattern emerged in session 6, where students with ChatGPT access scored higher on a procedurally complex quiz ($d=1.24$). Six out of seven sessions showed no meaningful between-group difference. The finding is interpreted as a conditional association at a point of accumulated procedural demand, not as evidence of a general AI advantage.

These findings suggest that the relationship between AI access and student performance may depend on when AI is made available and what kind of task students face. For course design, these results suggest that teachers could identify which sessions involve the highest procedural demands and plan AI access specifically for those sessions, while maintaining traditional instruction for conceptual content. This approach treats AI as a targeted support tool, not a permanent feature of every session. AI may help more in complex sessions. Whether this pattern holds in other courses and institutions needs to be tested in future research.

ACKNOWLEDGMENTS

The authors thank Adnane Afquire, Deputy Director in Charge of Pedagogical Affairs at the Higher International Institute of Tourism, Tangier, Morocco for administrative support and permission to conduct this research. The first author also thanks Laya Aziz for her constant support. The authors are also grateful to the participating students for their cooperation and engagement throughout the study.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Fatima Ezzahra Kabba	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	
Zouhair Ejbari	✓			✓						✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

Students were informed about the purpose of the study and consented to the use of their anonymized quiz data for research purposes. No personally identifiable information is reported in this manuscript. The study was conducted in compliance with applicable institutional standards and the Helsinki Declaration principles for research involving human participants.

ETHICAL APPROVAL

Students were informed about the study and consented to the use of anonymized quiz data. The study was conducted with institutional knowledge and approval. Quiz scores did not influence final grades.

DATA AVAILABILITY

The data that supports the findings of this study are available from the corresponding author, [FEK], upon reasonable request.

REFERENCES

[1] R. Michel-Villarreal, E. Vilalta-Perdomo, D. E. Salinas-Navarro, R. Thierry-Aguilera, and F. S. Gerardou, "Challenges and opportunities of generative AI for higher education as explained by ChatGPT," *Education Sciences*, vol. 13, no. 9, p. 856, Aug. 2023, doi: 10.3390/educsci13090856.

[2] H. Crompton and D. Burke, "The educational affordances and challenges of ChatGPT: state of the field," *TechTrends*, vol. 68, no. 2, pp. 380–392, Mar. 2024, doi: 10.1007/s11528-024-00939-0.

[3] A. Zeb, R. Ullah, and R. Karim, "Exploring the role of ChatGPT in higher education: opportunities, challenges and ethical considerations," *The International Journal of Information and Learning Technology*, vol. 41, no. 1, pp. 99–111, Jan. 2024, doi: 10.1108/IJILT-04-2023-0046.

[4] R. G. Luciano, "The dual impact of ChatGPT on learning and ethics among bachelor of science in information technology (BSIT) students," *International Journal of Learning, Teaching and Educational Research*, vol. 23, no. 12, pp. 78–95, Dec. 2024, doi: 10.26803/ijlter.23.12.5.

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

[6] J. Rudolph, S. Tan, and S. Tan, "ChatGPT: bullshit spewer or the end of traditional assessments in higher education?" *Journal of Applied Learning and Teaching*, vol. 6, no. 1, pp. 342–363, Jan. 2023, doi: 10.37074/jalt.2023.6.1.9.

[7] M. M. Asad and A. Ajaz, "Impact of ChatGPT and generative AI on lifelong learning and upskilling learners in higher education: unveiling the challenges and opportunities globally," *The International Journal of Information and Learning Technology*, vol. 41, no. 5, pp. 507–523, Nov. 2024, doi: 10.1108/IJILT-06-2024-0103.

[8] S. M. J. Uddin, A. Albert, A. Ovid, and A. Alsharef, "Leveraging ChatGPT to aid construction hazard recognition and support safety education and training," *Sustainability*, vol. 15, no. 9, p. 7121, Apr. 2023, doi: 10.3390/su15097121.

[9] O. A. Ojubanire, S. A. Olaleye, M. A. Marhraoui, M. W. Kamihanda, O. I. Oke, and O. A. Ojubanire, "Awareness, perception, and adoption of ChatGPT in African HEIs: a multi-dimensional analysis," *The Internet and Higher Education*, vol. 65, p. 100999, Apr. 2025, doi: 10.1016/j.iheduc.2025.100999.

[10] N. Eltaiba, S. Hosseini, and K. Okoye, "Benefits and impact of technology-enhanced learning applications in higher education in Middle East and North Africa: a systematic review," *Global Transitions*, vol. 7, pp. 350–374, 2025, doi: 10.1016/j.glt.2025.06.004.

[11] P. Bhaskar and C. K. Tiwari, "Charming or chilling? A comprehensive review of ChatGPT's in education sector," *The International Journal of Information and Learning Technology*, Mar. 2025, doi: 10.1108/IJILT-05-2024-0097.

[12] H. Al-Mughairi and P. Bhaskar, "Exploring the factors affecting the adoption AI techniques in higher education: insights from teachers' perspectives on ChatGPT," *Journal of Research in Innovative Teaching & Learning*, vol. 18, no. 2, pp. 232–247, Sep. 2025, doi: 10.1108/JRIT-09-2023-0129.

[13] D. Ravšelj *et al.*, "Higher education students' perceptions of ChatGPT: a global study of early reactions," *PLoS ONE*, vol. 20, no. 2, p. e0315011, Feb. 2025, doi: 10.1371/journal.pone.0315011.

[14] J. Hiebert and P. Lefevre, “Conceptual and procedural knowledge in mathematics: an introductory analysis,” in *Conceptual and Procedural Knowledge: The Case of Mathematics*, 1st ed., J. Hiebert, Ed., New York, NY: Routledge, 1986, pp. 1–27.

[15] J. R. Star, “Reconceptualizing procedural knowledge,” *Journal for Research in Mathematics Education*, vol. 36, no. 5, pp. 404–411, 2005.

[16] L. S. Vygotsky, *Mind in society: the development of higher psychological processes*. Cambridge, MA: Harvard University Press, 1978, doi: 10.2307/j.ctvjf9vz4.

[17] M. Prince, “Does active learning work? A review of the research,” *Journal of Engineering Education*, vol. 93, no. 3, pp. 223–231, Jul. 2004, doi: 10.1002/j.2168-9830.2004.tb00809.x.

[18] M. Tran, C. Balasooriya, C. Semmler, and J. Rhee, “Generative artificial intelligence: the ‘more knowledgeable other’ in a social constructivist framework of medical education,” *npj Digital Medicine*, vol. 8, no. 1, p. 430, Jul. 2025, doi: 10.1038/s41746-025-01823-8.

[19] P. Zhang and G. Tur, “A systematic review of ChatGPT use in K-12 education,” *European Journal of Education*, vol. 59, no. 2, p. e12599, Jun. 2024, doi: 10.1111/ejed.12599.

[20] S. W. Raudenbush and A. S. Bryk, *Hierarchical linear models: applications and data analysis methods*, 2nd ed. Thousand Oaks, CA: SAGE Publications, Inc., 2002.

[21] J. Cohen, *Statistical power analysis for the behavioral sciences*, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.

[22] S. Alneyadi and Y. Wardat, “ChatGPT: Revolutionizing student achievement in the electronic magnetism unit for eleventh-grade students in Emirates schools,” *Contemporary Educational Technology*, vol. 15, no. 4, p. ep448, Oct. 2023, doi: 10.30935/cedtech/13417.

[23] K. M. Selem, E. G. Saghier, S. H. A. Khreis, and C. C. Tan, “ChatGPT-related risk patterns and students’ creative thinking toward tourism statistics course: pretest and posttest quasi-experimentation,” *Journal of Hospitality & Tourism Education*, vol. 37, no. 3, pp. 313–328, Jul. 2025, doi: 10.1080/10963758.2025.2456638.

[24] T.-C. Yang, Y.-C. Hsu, and J.-Y. Wu, “The effectiveness of ChatGPT in assisting high school students in programming learning: evidence from a quasi-experimental research,” *Interactive Learning Environments*, vol. 33, no. 6, pp. 3726–3743, Jul. 2025, doi: 10.1080/10494820.2025.2450659.

[25] Y. Xing, “Exploring the use of ChatGPT in learning and instructing statistics and data analytics,” *Teaching Statistics*, vol. 46, no. 2, pp. 95–104, Apr. 2024, doi: 10.1111/test.12367.

[26] N. Alnaim *et al.*, “Effectiveness of ChatGPT in remote learning environments: an empirical study with medical students in Saudi Arabia,” *Nutrition and Health*, vol. 31, no. 3, pp. 1035–1046, Jul. 2025, doi: 10.1177/02601060241273596.

[27] W. Zhou and Y. Kim, “Innovative music education: an empirical assessment of ChatGPT-4’s impact on student learning experiences,” *Education and Information Technologies*, vol. 29, no. 16, pp. 20855–20881, Nov. 2024, doi: 10.1007/s10639-024-12705-z.

[28] J. Schwarz, “The use of generative AI in statistical data analysis and its impact on teaching statistics at universities of applied sciences,” *Teaching Statistics*, vol. 47, no. 2, pp. 118–128, May 2025, doi: 10.1111/test.12398.

[29] D. Thüs, S. Malone, and R. Brünken, “Exploring generative AI in higher education: a RAG system to enhance student engagement with scientific literature,” *Frontiers in Psychology*, vol. 15, p. 1474892, Oct. 2024, doi: 10.3389/fpsyg.2024.1474892.

[30] N. Moussa, R. Bentoumi, and T. Saali, “Promoting student success with neutrosophic sets: artificial intelligence and student engagement in higher education context,” *International Journal of Neutrosophic Science*, vol. 23, no. 1, pp. 238–248, 2024, doi: 10.54216/IJNS.230121.

[31] L. P. Patac and A. V. Patac, “Using ChatGPT for academic support: managing cognitive load and enhancing learning efficiency – a phenomenological approach,” *Social Sciences & Humanities Open*, vol. 11, p. 101301, 2025, doi: 10.1016/j.ssaho.2025.101301.

[32] D. Kaya and S. Yavuz, “Can generative AI and ChatGPT break human supremacy in mathematics and reshape competence in cognitive-demanding problem-solving tasks?” *Journal of Intelligence*, vol. 13, no. 4, p. 43, Apr. 2025, doi: 10.3390/jintelligence13040043.

[33] Z. Liu, H. Zuo, and Y. Lu, “The impact of ChatGPT on students’ academic achievement: a meta-analysis,” *Journal of Computer Assisted Learning*, vol. 41, no. 4, p. e70096, Aug. 2025, doi: 10.1111/jcal.70096.

APPENDIX

Assessment design, instructional context, and supporting materials

This appendix provides supporting information on the assessment design, instructional conditions, and quiz materials used in the study.

Instructional context and use of digital tools

Digital tools, including generative text-based applications, were permitted only during learning activities in designated classes and were not allowed during assessment. During learning activities, students in the AI-permitted condition were allowed to ask conceptual questions such as definitions and interpretations, request clarification of procedures discussed in class, compare explanations with lecture slides, and reformulate content in their own words. During quizzes, no digital tools or external assistance were permitted. Quizzes were completed individually under standard classroom supervision and administered at the end of each session.

Overview of quizzes and learning objectives

Quiz 1: role of statistics (conceptual). Quiz 2: statistical vocabulary (conceptual). Quiz 3: types of variables (conceptual). Quiz 4: frequencies (procedural/application). Quiz 5: means (procedural/application). Quiz 6: harmonic mean and mode (procedural/application). Quiz 7: median and graphs (application).

Pedagogical signals from assessment performance

Across sessions, certain concepts and procedures consistently generated student difficulty. In quiz 1, students often interpreted descriptive statistics as data rather than as methods. In quiz 2, defining statistical populations beyond the common understanding of 'people' proved challenging. In quiz 3, justifying the classification of ordinal variables in satisfaction scales was a recurring source of confusion. Quiz 4 revealed difficulties in computing and interpreting frequency distributions correctly. In quiz 5, calculating the geometric mean presented persistent challenges. In quiz 7, selecting appropriate graphical representations alongside determining the median required judgment that students found demanding.

Verbatim assessment items for quiz 6

To ensure transparency regarding the assessment associated with session 6, this appendix reports the exact wording and structure of the items administered in quiz 6. These items were identical across all classes and conditions.

– Item 1: harmonic mean (discrete data with frequencies)

Calculate the harmonic mean for the following data: speed (km/h): 20, 30, 40, 50, 60 with frequency: 4, 6, 8, 2, 1. Response options: A) 31.98 km/h, B) 35.4 km/h, C) 36.2 km/h, D) I do not know. This item required correct application of the harmonic mean formula with frequency weights.

– Item 2: mode (discrete variable)

Determine the mode for the following data: Rassi (2 children), Belabasse (5), Azeroual (6), Bolif (5). Response options: A) Rassi, B) Belabasse, C) Azeroual, D) I do not know. This item assessed conceptual understanding of the mode as the most frequently occurring value.

– Item 3: mode for grouped continuous data (unequal class widths)

Determine the mode for the following continuous variable: Salary interval [0; 3000) Freq=10,000 Width=3,000 Density=3.33; [3000; 6000) Freq=25,000 Width=3,000 Density=8.33; [6000; 7000) Freq=30,000 Width=1,000 Density=30.00; [7000; 9000) Freq=20,000 Width=2,000 Density=10.00; [9000; 12000) Freq=15,000 Width=3,000 Density=5.00; [12000; 15000) Freq=5,000 Width=3,000 Density=1.67. Response options: A) 3000.66, B) 5390.35, C) 6333.33, D) I do not know.

Item 3 required identification of the modal class based on density rather than raw frequency, followed by application of the interpolation formula for grouped continuous data. Together, the three items required multi-step computation, correct formula selection, and procedural sequencing, reflecting a substantially higher cognitive and computational load than assessments used in earlier sessions.

Ethical and reporting considerations

No raw student responses are disclosed and no individual interaction logs were collected. All results are reported at the group level to preserve student confidentiality. ChatGPT was available to students in the AI-permitted classes only during learning activities and was not permitted during quizzes in any class.

BIOGRAPHIES OF AUTHORS

Fatima Ezzahra Kabba holds a Ph.D. in Economics and Management (2024) and a master's degree in Finance. She is a lecturer at the Higher International Institute of Tourism, Morocco, where she teaches economics and management, and an affiliated researcher with the Faculty of Legal, Economic and Social Sciences of Tangier, Abdelmalek Essaâdi University. She also holds the Google Data Analytics Professional Certificate. Her research interests focus on the impact of artificial intelligence on learning and data-driven decision-making in education. She can be contacted at: kabba.fatima.ezzahra@gmail.com.

Zouhair Ejbari holds a Ph.D. in Economics and Management. He is a senior professor at the Faculty of Legal, Economic and Social Sciences of Tangier, Abdelmalek Essaâdi University, Morocco. His teaching and research focus on public management and higher education. His research interests include governance, public policy, and data-driven analysis in educational and institutional contexts. He can be contacted at email: zejbari@uae.ac.ma.