Revista Española de Educación Médica

UNIVERSIDAD DE MURCIA

Declared preferences for virtual tutoring styles in two Kinesiology curriculum cohorts: an exploratory cross-sectional study.

Declared preferences for virtual tutoring styles in two curricular cohorts of Kinesiology students: an exploratory cross-sectional study .

Italo Campos Montenegro [1]

[1] Kinesiology Program, Faculty of Life Sciences, Viña del Mar University, Viña del Mar, Chile, italo.campos@uvm.cl, https://orcid.org/0009-0007-0325-3344

* Correspondence: italo.campos@uvm.cl

Received: 24/6/26; Accepted: 28/7/26; Published: 30/7/26

Summary.

Aim : To compare the reported preference for virtual tutoring styles between two curricular cohorts of Kinesiology students and to describe differences in perceived usefulness, study strategies, and functional orientation of tutor use. **Methods** : A quantitative, observational, and cross-sectional study was conducted with 6th and 8th semester Kinesiology students at the University of Viña del Mar. Students had free access to two virtual tutors—one structured-explanatory and the other reflective-Socratic—without assignment, standardized exposure, or objective recording of interactions. The main variable was the stated preference for the reflective-Socratic tutor. Secondary outcomes were obtained using an ad hoc questionnaire and interpreted as self-reported contextual indicators . The difference between academic cohorts in preference for the reflective-Socratic tutor was estimated using Fisher's exact test and Firth's penalized logistic regression, adjusted for frequency of use and prior experience with online tutors. **Results** : 44 of 70 invited students participated. Preference for the reflective-Socratic tutor was reported by 10% of 6th-semester students and 70.8% of 8th-semester students. The association was significant in the unadjusted analysis and remained significant after adjustment using Firth's penalized logistic regression. Perceived usefulness was high in both groups. Secondary results showed different descriptive patterns in reflection, justification, and the creation of summaries or outlines, but none of the eight comparisons retained statistical significance after Holm's adjustment. **Conclusions** : Marked differences were observed in the reported preference for virtual tutoring styles between the two curricular cohorts studied. Because each cohort corresponded to a different subject and had different contextualized tutors, the results do not allow these differences to be attributed solely to the educational level. The observed patterns allow us to formulate hypotheses about the relationship between tutoring configuration, curricular context, task characteristics, and expected level of autonomy, which should be evaluated using equivalent exposure designs and objective educational outcomes.

Keywords: virtual tutor; artificial intelligence; ChatGPT; Socratic tutoring; clinical reasoning; health education; higher education.

Abstract

Objective: To compare the reported preference for virtual tutoring styles between two curricular cohorts of Kinesiology students and to describe differences in perceived usefulness, study strategies, and functional orientation of tutor use. **Methods:** A quantitative, observational, cross-sectional study was conducted with 6th and 8th semester Kinesiology students at Universidad Viña del Mar. Students had open access to two virtual tutors—one structured-explanatory and the other reflective-

Socratic—without assignment, standardized exposure, or objective recording of interactions. The primary variable was the reported preference for the reflective-Socratic tutor. Secondary outcomes were obtained through an ad hoc questionnaire and interpreted as contextual indicators of self-reporting. The difference between curricular cohorts in preference for the reflective-Socratic tutor was estimated using Fisher's exact test and Firth's penalized logistic regression, adjusted for frequency of use and prior experience with virtual tutors. **Results:** 44 of 70 invited students participated. Preference for the reflective-Socratic tutor was reported by 10% of sixth-semester students and by 70.8% of eighth-semester students. The association was significant in the unadjusted analysis and remained so after adjustment using Firth's penalized logistic regression. Perceived usefulness was high in both groups. Secondary outcomes showed different descriptive patterns in reflection, justification, and the creation of summaries or outlines, but none of the eight comparisons retained statistical significance after Holm's adjustment. **Conclusions** : Marked differences were observed in the reported preference for online tutoring styles between the two curricular cohorts studied. Because each cohort corresponded to a different subject and different contextualized tutors, the results do not allow these differences to be attributed solely to the educational level. The observed patterns allow us to formulate hypotheses about the relationship between tutoring configuration, curricular context, task characteristics, and expected degree of autonomy, which should be evaluated using designs with equivalent exposure and objective educational outcomes.

Keywords: virtual tutor; artificial intelligence; ChatGPT; Socratic tutoring; clinical reasoning; health professions education; higher education.

1. Introduction

The integration of conversational virtual tutors in higher education has grown steadily, establishing itself as an emerging strategy to support learning in diverse contexts. These systems—ranging from educational chatbots to agents based on large-scale language models, such as ChatGPT—have been incorporated into explanatory, support, evaluative, and reflective functions. The current discussion has begun to shift from technological availability to the quality of pedagogical mediation: how these tutors interact with students, what type of cognitive engagement they promote, and under what conditions their design is most suitable for different learning profiles (1, 2) . A critical dimension in this discussion is the pedagogical style with which the virtual tutor structures the interaction. Generally speaking, one can distinguish between structured-explanatory approaches, focused on organizing content, providing guided explanations, sequencing information, and supporting the synthesis of material, and reflective-Socratic approaches, focused on guiding questions, contrasting assumptions, justifying answers, and stimulating autonomous reasoning. This distinction does not represent an absolute opposition between providing answers and formulating questions, but rather a difference in the predominant pedagogical emphasis. While some designs reduce initial learning friction through structure, clarity, and guidance, others seek to increase active student participation through reflection, argumentation, and metacognition (3, 4).

Recent evidence shows that conversational tutors can influence multiple dimensions of learning, including performance, motivation, cognitive load, engagement, technological acceptance, and critical reasoning. However, their effectiveness depends on how they are pedagogically integrated and how their interactions support self-regulation, metacognitive monitoring, and feedback. From this perspective, the value of the virtual tutor lies not only in responding quickly or being constantly available, but also in organizing, challenging, and supporting the learning process in a way that is consistent with the educational objectives (5, 6). This distinction is especially relevant because frequent use of a virtual tutor does not necessarily equate to pedagogically effective use. When the system delivers final products without promoting elaboration, verification, or justification, it can facilitate immediate performance while simultaneously limiting deep conceptual understanding. Recent studies have shown that certain AI supports can improve performance during a task or reduce frustration without producing equivalent learning gains, and that the use of generative AI without

pedagogical safeguards can negatively affect learning when it displaces the student's cognitive effort. In this regard, over-reliance on conversational AI systems has been identified as a risk to cognitive skills such as critical thinking, autonomous judgment and decision making (7-9) .

In this context, student perceptions provide relevant signals for educational design, although they require careful interpretation. Satisfaction, perceived usefulness, or stated preference for a tutoring style do not, in themselves, demonstrate educational effectiveness or deep learning. A tutor may be valued for their clarity, speed, or availability, and still promote interactions of low cognitive demand. Similarly, a reflective tutor may stimulate reasoning and autonomy, but increase the cognitive load if they are not suited to the student's educational level, task type, or prior mastery. Therefore, studying student preferences is relevant when they are interpreted as indicators of perceived pedagogical adequacy and not as direct evidence of the superiority of one style over another (2, 4). This limitation has also been applied to Kinesiology students exposed to a GPT-based tutor, where the associations between self-reported usage conditions and subjective experience were interpreted as exploratory and contextual, without inferring causality or effectiveness on academic performance (10).

This discussion is particularly relevant in health professions, where professional training demands progression from understanding and organizing content to integrating information, making decisions, and justifying clinical or therapeutic actions. In these types of learning paths, the stage of training can modify the expected pedagogical role of the online tutor. Students at intermediate stages may require more structure, synthesis, and guidance to organize knowledge, while advanced students may place greater value on reflection, comparing alternatives, and justifying decisions. Despite the growth of the literature on educational artificial intelligence, evidence on AI interventions in health professions remains limited, especially when considering complex educational outcomes, training pathways, and pedagogical designs tailored to the progressive development of professional reasoning (11). Along these same lines, it has been noted that generative AI can support critical thinking in medical education, but it can also strain cognitive autonomy when its use displaces the student's own processes of analysis, verification, and judgment (12).

In this context, this cross-sectional study compares the reported preference for virtual tutoring styles between two cohorts of Kinesiology students and integrates indicators of perceived usefulness, study strategies, and functional orientation of use. The cohorts corresponded to the 6th and 8th semesters and used tutors with a common pedagogical approach, although contextualized to the content and demands of their respective courses. Therefore, the purpose of the study is to describe situated differences between the two curricular contexts and explore the hypothesis that tutorial mediation needs may vary throughout the course of study, without attributing these differences exclusively to the level of study .

2. Methods

2.1 Study design

A quantitative, cross-sectional, observational, and associational study was conducted using a self-administered survey of undergraduate Kinesiology students exposed to the use of virtual tutors for academic support. The study was carried out in a naturalistic educational implementation context, without random assignment or experimental control of exposure. The main factor of comparison was the curricular cohort, defined by enrollment in the 6th or 8th semester and the corresponding course. Because each cohort used tutors configured with their own content, activities, and curricular demands, the semester, course, and specific tutor implementation were completely overlapping. Consequently, the results are interpreted as differences associated with two curricular contexts and not as an independent estimate of the effect of the educational level. The study was designed to

describe preferences, perceptions, and declared usage patterns, without estimating effects on learning, academic performance, clinical reasoning, cognitive autonomy, or educational transfer.

2.2 Context, population and sample

The study population consisted of 70 regular students enrolled in the Kinesiology program at Universidad Viña del Mar, belonging to two academic cohorts: one in the 6th semester and the other in the 8th semester. Each cohort corresponded to a different course in which virtual tutors were incorporated as a resource to support independent learning . A non-probability sampling method was used, with all eligible students invited to participate. Inclusion criteria were: being enrolled in the 6th or 8th semester of the Kinesiology program, having had access to the virtual tutors provided for the corresponding course, accepting the digital informed consent form, and completing the survey sufficiently for analysis. Records with more than 30% unanswered items or with evident, unrecoverable inconsistencies were excluded. The final sample comprised 44 students, representing 62.9% of the invited population. Given the exploratory and contextual nature of the study, a formal sample size calculation aimed at estimating population effects was not performed. The sample consists of eligible students who agreed to participate and completed the survey sufficiently for analysis. Consequently, the results should be interpreted as associational evidence situated within a specific cohort, without claiming statistical representativeness of other health programs, institutions, or disciplines.

2.3 Description of virtual tutors

In each course cohort, two conversational virtual tutors were created using the "Create GPT" function of ChatGPT Plus. The tutors were configured according to the learning outcomes, content, activities, and official bibliography of each subject. Each course included one structured-explanatory tutor and one reflective-Socratic tutor, both available for voluntary and free use. Students could access either tutor and independently decide which to use, how often, and for what academic purpose. No usage sequence, minimum exposure time, or equivalent number of interactions with each tutor was assigned. The number of sessions, interaction duration, the tutor used on each occasion, and the proportion of use corresponding to each style were also not recorded. Free access allowed the tutors to be integrated under naturalistic conditions of academic use, but it did not ensure that all participants used or compared both styles under equivalent conditions. Consequently, the stated preference may reflect a combination of differential familiarity, self-selection, intensity of use, particular academic needs, and prior affinity with a certain type of interaction. For this reason, it is not interpreted as an experimental measure of preference nor as evidence of the pedagogical superiority of one tutor over another.

Although both curricular cohorts shared the same general pedagogical architecture—a structured-explanatory tutor and a reflective-Socratic tutor—the tutors did not provide identical interventions. Each pair was configured with the learning outcomes, content, activities, and bibliography specific to their respective subject. Therefore, the comparison should be interpreted as a comparison between two cohorts exposed to tutors with equivalent general pedagogical orientations, but contextualized to different subjects and curricular demands. This overlap makes it impossible to isolate the independent effect of the educational level with respect to the subject studied, the complexity of the tasks, the curriculum, and the specific implementation characteristics of each tutor

In operational terms, the structured-explanatory tutor was configured to prioritize the progressive organization of content, the delivery of direct and understandable explanations, the breakdown of problems or tasks into steps, and support in developing summaries, outlines, and study strategies. Its pedagogical approach aimed to reduce the initial difficulty of accessing the content through explicit structure, sequencing, and academic guidance. The reflective-Socratic tutor was configured to guide the interaction through guiding questions, follow-up questions, contrasting assumptions, and requesting justification of answers. Its pedagogical approach aimed to have the

student make their reasoning explicit, examine alternatives, and review their own decisions before moving forward in the interaction. It also incorporated questions designed to encourage metacognitive reflection on the process used to arrive at the answer.

To promote reproducibility, the core instructions used in configuring both tutoring styles are presented in Supplementary Material 1. This document includes the pedagogical purpose, interaction rules, expected type of support, how to structure responses, and common restrictions. The instructions correspond to the configurations actually used during implementation and are presented by differentiating the shared pedagogical architecture from the specific curriculum of each subject. Both tutors were configured with common restrictions aimed at keeping their use within the academic framework of the subject. It was established that they should respond based on the learning outcomes, content, activities, and bibliography incorporated into their configuration. Furthermore, they should not replace the instructor, issue medical diagnoses, grade assignments, complete formal assessments, provide psychological support, or answer questions outside the defined curriculum. Conversations between students and virtual tutors were neither collected nor analyzed.

2.4 Instrument and variables

Data collection was conducted using a self-administered structured survey in Microsoft Forms. The instrument was developed specifically for this study by the researcher, based on the research question, the study objectives, and the dimensions considered relevant for characterizing the user experience of virtual tutors in the participating courses. The questionnaire was designed for this specific educational implementation and not as a psychometric scale intended to measure generalizable constructs. The instrument was organized into five domains: general data; prior experience and frequency of use of online tutors; stated preference for tutoring style; perceived usefulness, perceived support for performance, and tutor-assisted study strategies; and functional orientation of use. This last dimension included items related to structured explanation, guidance toward the correct answer, reflective questions, justification of decisions or answers, and the creation of summaries or outlines based on the tutor's responses.

The survey included closed-ended multiple-choice questions, five-point Likert-type items—ranging from 1 = strongly disagree to 5 = strongly agree—and two open-ended questions designed to gather the most valued aspects of the online tutor and suggestions for improvement. The primary variable was obtained through a closed-ended question in which participants stated their preference for either a structured-explanatory tutor or a reflective-Socratic tutor. Consequently, this variable represents a specific, contextually declared choice, not a psychometric scale or the measurement of a latent construct. The primary comparison variable was the academic cohort, coded as 6th or 8th semester. Frequency of use and prior experience with online tutors were considered covariates. Frequency of use was recorded through an ordinal self-report question with the categories “Never,” “Occasionally,” “Weekly,” “Several times a week,” and “Daily.” This variable represented the overall frequency of virtual tutor use during the implementation period and did not allow us to identify which tutor was used, how many times each was accessed, the duration of the interactions, or whether the two styles were actually compared. The “Never” category identified students who, despite having access, reported not using the tutors during the study period. Consequently, frequency of use was considered a self-reported contextual covariate and not an objective or tutor-specific measure of exposure. The reported preference should be interpreted within these uncontrolled exposure conditions. Its association with the curriculum cohort does not allow us to determine whether the observed differences resulted from an equivalent comparison between tutoring styles .

For secondary outcomes, three exploratory composite scores were constructed from predefined item groupings: perceived usefulness, tutor-assisted study strategies, and perceived performance support. Indicators related to the tutor's functional orientation were analyzed individually. The

internal consistency of the three composite scores was estimated using Cronbach's alpha in the studied sample. These coefficients were used solely to describe the degree of internal relationship between items and not as evidence of content, construct, or criterion validity. The instrument was not subjected to formal psychometric validation or factor analysis. Therefore, its results should be interpreted as contextual and exploratory indicators of preference, perception, and reported use. The complete wording of the items, their response formats, and the composition of the exploratory composite scores are presented in Supplementary Material 2.

2.5 Data collection procedure

The educational implementation of virtual tutors was carried out as part of the regular academic support for the courses involved. Data collection for research purposes was conducted after approval by the Scientific Ethics Committee of the University of Viña del Mar. In January 2026, students received an institutional invitation with a link to the survey on Microsoft Forms. The use of the tutors was voluntary and was not part of a mandatory activity, experimental sequence, or pre-established pedagogical plan. Students were not required to use both tutors before completing the survey . Before responding, participants accessed a digital informed consent form that explained the study's objectives, the voluntary nature of participation, data confidentiality, and the absence of academic consequences associated with participating or not participating. Only those who accepted the informed consent form were allowed to proceed with the survey. The data were exported from Microsoft Forms to an anonymized database for analysis.

2.6 Data Management and Security

Data were treated anonymously or pseudonymized for analysis. No direct identifiers or records of conversations with online tutors were collected. Open-ended responses were reviewed to detect and remove potential identifying elements before qualitative analysis. The database was stored with access restricted to the principal investigator, and results were reported in aggregate form.

2.7 Data Analysis

A descriptive analysis of the sample was performed using frequencies, percentages, means, and standard deviations, depending on the nature of the variables. The association between academic cohort—8th versus 6th semester—and reported preference for the reflective-Socratic tutor was examined using Fisher's exact test. The odds ratio and its 95% confidence interval were reported. Additionally, a Firth penalized logistic regression model was fitted, using preference for the reflective-Socratic tutor as the dependent variable. The model incorporated academic cohort, self-reported overall frequency of use, and prior experience with online tutors as predictors. Penalized regression was used due to the sample size and the presence of small cells in the main contingency table, conditions under which conventional logistic regression can produce unstable estimates. The model was deliberately kept parsimonious and included only the predefined covariates considered relevant for interpreting the main association. Confidence intervals were estimated using a penalized likelihood profile. The subject matter and the specific tutor configuration were not included as covariates because they were entirely consistent with the curriculum cohort, and their effects could not be estimated independently. Therefore, the cohort estimator was interpreted as a contextual comparison between the two groups studied, rather than as the isolated effect of the educational level. The frequency of use corresponded to a general self-reported measure, not a specific measure of exposure to each tutor. Consequently, its inclusion in the model did not allow for controlling for differential familiarity, the intensity of interactions, or an effective comparison between the two tutoring styles. The adjusted estimators were interpreted considering the exploratory nature of the study, the sample size, and the width of the confidence intervals.

As a sensitivity analysis, the model was replicated excluding the two participants who reported not using the virtual tutors during the implementation period. The aim was to determine whether the

main difference between cohorts depended on the inclusion of these cases. This analysis was not designed to correct for the lack of equivalent exposure, as objective or specific usage records were not available for the participants who reported using the tutors. Internal consistency was estimated using Cronbach's alpha for the three exploratory composite scores—perceived usefulness, tutor-assisted study strategies, and perceived performance support. The analysis was supplemented with corrected item-total correlations and with the calculation of alpha if each item was removed. For the two-item composite score, the correlation between the two items was also examined. These analyses were used to describe the internal behavior of the scores in the studied sample and not as evidence of validity, unidimensionality, or psychometric strength of the instrument. Factor analyses were not performed, as the study was not designed as an instrumental validation process. The three composite scores and the five individual functional indicators were compared between cohorts using the two-tailed Mann-Whitney U test. For each comparison, Cliff's delta was reported as a nonparametric effect size. Positive values indicated higher scores in the 8th-semester cohort, and negative values indicated higher scores in the 6th-semester cohort. The eight secondary comparisons were considered a single analytical family and were corrected using Holm's sequential procedure to control for familial error arising from multiplicity. Unadjusted and adjusted p-values were reported. The primary outcome of tutoring style preference was not included in this correction because it corresponded to the primary variable and was analyzed using a differentiated hypothesis. Secondary outcomes were interpreted considering the adjusted values, the direction and magnitude of Cliff's delta, and the exploratory nature of the study.

The association between overall self-reported frequency of use and perceived usefulness was also examined using Spearman's rho coefficient. Due to the cross-sectional design, this relationship was interpreted without establishing a causal direction.

Open-ended responses were examined using thematic content analysis with a deductive-inductive approach. The procedure included reading all responses, constructing specific thematic categories, coding the content, and synthesizing the main themes. Qualitative results were used to contextualize the quantitative findings and not as independent evidence of educational outcomes. All tests were two-tailed, and a significance level of 0.05 was used. Statistical analyses were performed using Python version 3.13. The code used for data processing and analysis, along with an anonymized version of the analytical matrix, is available upon reasonable request to the corresponding author, in accordance with the ethical and confidentiality commitments established for the study.

2.8 Ethical considerations

The study was approved by the Scientific Ethics Committee for Research in Humans, Animals, and All Other Living Beings of the University of Viña del Mar, through Ethical Evaluation Report No. 09-26, issued on January 9, 2026, corresponding to the project "Preferences and Effects of the Use of Academic and Socratic Virtual Tutors in Kinesiology Students at the University of Viña del Mar." The resolution established that the proposed procedures did not present any ethical concerns . All participants provided digital informed consent before answering the survey. Participation was voluntary, anonymous, and without academic consequences. Data confidentiality was maintained, and the results are presented only in aggregate form.

3. Results

3.1 Participants and usage pattern

Forty-four students responded out of a total of 70 invited, representing an overall response rate of 62.9%. The sample consisted of 20 sixth-semester students and 24 eighth-semester students. The average age was 24.8 years (SD = 4.0; n = 43), with one implausible record entered as "2022" being considered missing data. The gender distribution was balanced in the total sample, with 22 women and 22 men. Prior experience with online tutors was common in both groups: 17 out of 20 sixth-

semester students and 18 out of 24 eighth-semester students reported having previously used this type of tool. The overall self-reported frequency of use showed a heterogeneous pattern. In the 6th-semester cohort, occasional use several times a week predominated, with no students reporting daily use. In the 8th-semester cohort, occasional use also predominated; two students reported not using the tutors, and three reported daily use. This variable did not allow us to identify which of the two tutors was used or the specific intensity of exposure to each style. The two cases classified as "Never" were retained in the sample description and the main analysis, since they belonged to the eligible cohort, had access to the resource, and answered the preference question. Their influence was examined through a sensitivity analysis . Age is reported as n=43 due to an implausible record entered as "2022," treated as missing data only for this variable. The frequency corresponds to the overall self-reported use of virtual tutors and does not distinguish between the structured-explanatory tutor and the reflective-Socratic tutor.

Table 1. Sample characteristics and usage pattern.

Feature	6th semester (n=20)	8th semester (n=24)	Total (n=44)
Age, mean (SD)	24.1 (4.2) [n=19]	25.5 (3.8) [n=24]	24.8 (4.0) [n=43]
Women, n (%)	7 (35.0)	15 (62.5)	22 (50.0)
Men, n (%)	13 (65.0)	9 (37.5)	22 (50.0)
Previous experience with virtual tutors, n (%)	17 (85.0)	18 (75.0)	35 (79.5)
Overall self-reported frequency of use, n (%)	—	—	—
Never	0 (0,0)	2 (8,3)	2 (4,5)
Occasional	8 (40.0)	9 (37.5)	17 (38.6)
Weekly	5 (25.0)	5 (20.8)	10 (22.7)
Several times/week	7 (35.0)	5 (20.8)	12 (27,3)
Daily	0 (0,0)	3 (12.5)	3 (6,8)

3.2 Preference for virtual tutor style

A marked difference was observed in the reported preference for the reflective-Socratic tutor between the two academic cohorts. In the 6th-semester cohort, 2 of 20 students (10.0%) preferred this tutoring style, while in the 8th-semester cohort, 17 of 24 students (70.8%) reported this preference. The absolute difference between groups was 60.8 percentage points, and the ratio of proportions was 7.08. Fisher's exact test showed a statistically significant association between academic cohort and preference for the reflective-Socratic tutor ($p < 0.001$), with an odds ratio of 21.86 (95% CI: 3.97–120.31).

In Firth's penalized logistic regression model, adjusted for frequency of use and prior experience with online tutors, the difference between cohorts remained. Students in the 8th-semester cohort had higher odds of preferring the reflective-Socratic tutor than students in the 6th-semester cohort (aOR = 14.79; 95% CI: 3.64–85.67; $p < 0.001$). Frequency of use (aOR = 1.14; 95% CI: 0.57–2.37; $p = 0.716$) and prior experience with online tutors (aOR = 0.61; 95% CI: 0.09–4.03; $p = 0.609$) did not show statistically significant associations with reported preference. The cohort estimator represents the joint difference between the studied contexts and does not allow separating the educational level of the subject and the specific configuration of the tutors .

The sensitivity analysis, performed after excluding the two participants who reported not using the virtual tutors, showed a pattern consistent with the main analysis (aOR = 12.87; 95% CI: 3.11–75.30; $p < 0.001$). This result indicates that the observed difference between cohorts was not dependent on the inclusion of these two cases. However, the analysis does not allow for correcting for differential exposure among tutors, as objective and tutor-specific records of usage were not available for those who reported using them .

Table 2. Declared preference for virtual tutoring style and partnership models.

Variable	6th semester (n = 20)	8th semester (n = 24)	OR/aOR Firth (95% CI)	p
Preference for structured-explanatory tutor, n (%)	18 (90.0)	7 (29,2)	—	—
Preference for reflective-Socratic tutor, n (%)	2 (10,0)	17 (70.8)	OR = 21.86 (3.97–120.31)	<0.001
Absolute difference in preference for reflective-Socratic tutor	—	+60.8 percentage points	—	—
Ratio of proportions for preference for reflective-Socratic tutor	—	7.08	—	—
Firth-adjusted model: 8th semester cohort vs. 6th semester cohort	—	—	aOR = 14.79 (3.64–85.67)	<0.001
Frequency of use, by category	—	—	aOR = 1.14 (0.57–2.37)	0.716
Previous experience with online tutors, yes vs. no	—	—	aOR = 0.61 (0.09–4.03)	0.609
Sensitivity analysis without participants who stated "Never"	—	—	aOR = 12.87 (3.11–75.30)	<0.001

OR = unadjusted odds ratio; aOR = adjusted odds ratio; CI95% = 95% confidence interval. The unadjusted OR corresponds to the association between curriculum cohort and preference for reflective-Socratic tutor estimated from the contingency table. The adjusted model was estimated using Firth's penalized logistic regression and included curriculum cohort, overall self-reported frequency of use, and prior experience with online tutors . The confidence intervals for the penalized model correspond to the penalized likelihood profile. The sensitivity analysis excluded the two participants who reported not using the online tutor during the implementation

period. Due to complete overlap between cohort, subject, and tutor configuration, the estimator does not represent the independent effect of educational level. The adjustment for overall frequency of use and the sensitivity analysis do not control for specific exposure to each tutor.

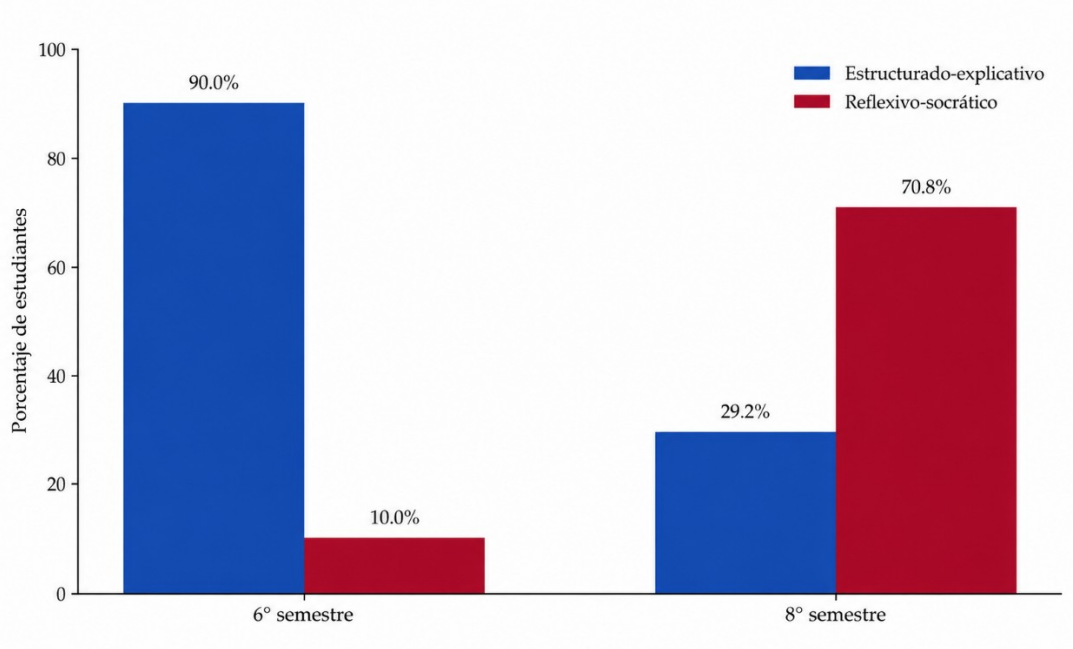


Figura 1. Preferencia declarada por estilo de tutoría virtual según cohorte curricular.

3.3 Perceived usefulness, study strategies and functional indicators

The three exploratory composite scores showed high values in both cohorts. The perceived usefulness scale, composed of six items, showed an internal consistency of $\alpha = 0.790$. Corrected item-total correlations ranged from 0.401 to 0.694, while alpha coefficients obtained after removing each item ranged from 0.720 to 0.792. Excluding one item did not produce a relevant improvement in the score's internal consistency. The study strategies score, also composed of six items, showed more modest internal consistency, with $\alpha = 0.683$. Corrected item-total correlations ranged from 0.282 to 0.658, and alpha coefficients obtained after removing each item ranged from 0.562 to 0.690. No single item alone explained the lower internal consistency observed. This behavior suggests that the lower homogeneity may be related to the diversity of strategies included in the scoring, among them reformulating concepts, summarizing, analyzing perspectives, reflecting on responses, comparing reasoning, and revising errors. For this reason, it was interpreted as a contextual exploratory grouping and not as a consolidated unidimensional scale.

Perceived performance support consisted of two items and showed an internal consistency of $\alpha = 0.867$. The correlation between the two items was $r = 0.770$. Due to the small number of items, this result was interpreted as a brief composite score rather than a consolidated psychometric scale. Taken together, these analyses describe the internal behavior of the scores in the studied sample, but do not constitute evidence of the instrument's validity or unidimensionality. The composite scores did not show statistically significant differences between cohorts. Perceived usefulness had means of 4.48 (SD = 0.50) in the 6th semester and 4.41 (SD = 0.67) in the 8th semester (unadjusted $p = 0.923$; Holm-adjusted $p = 1.000$; $\Delta = -0.02$). Tutor-assisted study strategies had means of 4.43 (SD = 0.56) and 4.19 (SD = 0.56), respectively (unadjusted $p = 0.179$; adjusted $p = 0.716$; $\Delta = -0.24$). Perceived performance support was also comparable between cohorts, with means of 4.25 (SD = 0.72) and 4.17 (SD = 0.88) (unadjusted $p = 0.912$; adjusted $p = 1.000$; $\Delta = -0.02$). In the individual functional indicators, both

cohorts showed high scores in detailed and structured explanations: 4.70 (SD = 0.57) in the 6th semester and 4.79 (SD = 0.51) in the 8th semester (unadjusted p = 0.524; adjusted p = 1.000; Δ = 0.08). Direct guidance to the correct answer showed means of 4.40 (SD = 0.88) and 4.75 (SD = 0.61), respectively (unadjusted p = 0.100; adjusted p = 0.501; Δ = 0.23). The 8th semester cohort showed higher scores on items related to questions that promote reflection—4.15 (SD = 0.99) versus 4.67 (SD = 0.70)—and justification of decisions or responses—4.45 (SD = 0.94) versus 4.83 (SD = 0.48). However, none of these differences remained statistically significant after Holm's adjustment: reflective questions, unadjusted p = 0.046, adjusted p = 0.324 and Δ = 0.30; justification of decisions or responses, unadjusted p = 0.047, adjusted p = 0.324 and Δ = 0.27.

The largest difference among the secondary outcomes was in the creation of summaries or outlines. The mean was 4.65 (SD = 0.59) in the 6th semester and 3.75 (SD = 1.29) in the 8th semester, with an effect size of Δ = −0.44. Although the unadjusted p-value was 0.007, the difference did not reach the conventional significance threshold after Holm's adjustment (adjusted p = 0.057). Consequently, none of the eight secondary comparisons between cohorts retained statistical significance after controlling for multiplicity. These results should be interpreted as exploratory patterns of self-reporting and not as confirmatory evidence of functional differences between cohorts. Overall self-reported frequency of use was positively associated with the perceived usefulness of the online tutor (Spearman's rho = 0.399; p = 0.007). Participants who reported more frequent use also reported greater perceived usefulness. Due to the cross-sectional design, this association does not allow us to determine whether a higher tutor rating favored its use, whether more frequent use increased the perception of usefulness, or whether both variables were influenced by unmeasured personal or academic factors.

Table 3. Exploratory composite scores and self-reported functional indicators.

Result	6th semester, average (DE)	8th semester, average (DE)	unadjusted p	p adjusted by Holm	Cliff's Δ
Perceived utility, α = 0.79	4.48 (0.50)	4.41 (0.67)	0.923	1,000	−0.02
Study strategies, α = 0.68	4.43 (0.56)	4.19 (0.56)	0.179	0.716	−0.24
Perceived support for performance, α = 0.87	4.25 (0.72)	4.17 (0.88)	0.912	1,000	−0.02
Detailed and structured explanations	4.70 (0.57)	4.79 (0.51)	0.524	1,000	0.08
Direct guide to the correct answer	4.40 (0.88)	4.75 (0.61)	0.100	0.501	0.23
Questions that encourage reflection	4.15 (0.99)	4.67 (0.70)	0.046	0.324	0.30
Justification of decisions/responses	4.45 (0.94)	4.83 (0.48)	0.047	0.324	0.27
Preparation of summaries/outlines	4.65 (0.59)	3.75 (1.29)	0.007	0.057	−0.44

Between-cohort comparisons were performed using the two-tailed Mann-Whitney U test. The eight secondary comparisons were corrected using Holm's sequential procedure. Cliff's delta (Δ) was reported as a nonparametric effect size; positive values indicate higher scores in the 8th-semester cohort, and negative values indicate higher scores in the 6th-semester cohort. Perceived usefulness and study strategies each consisted of six items, while perceived performance support consisted of two items. Alpha coefficients describe the internal consistency observed in this sample and do not constitute evidence of validity, unidimensionality, or psychometric strength of the scores. For the perceived performance support score, the correlation between the two items was r = 0.770. No secondary comparison retained statistical significance after adjustment for multiplicity.

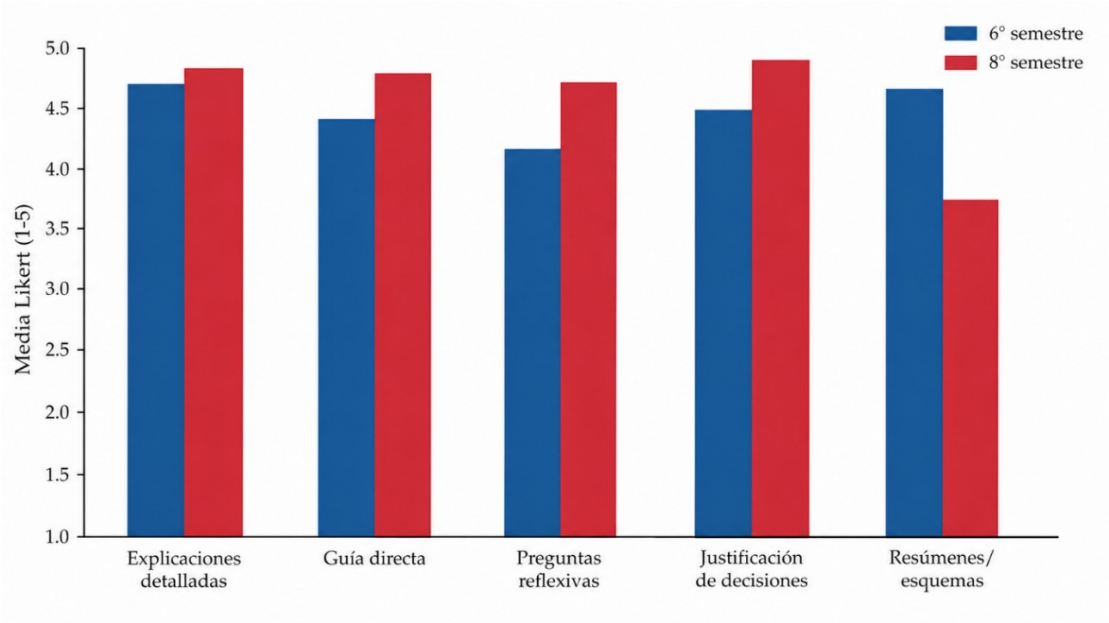


Figura 2. Perfil descriptivo de los indicadores funcionales autorreportados.

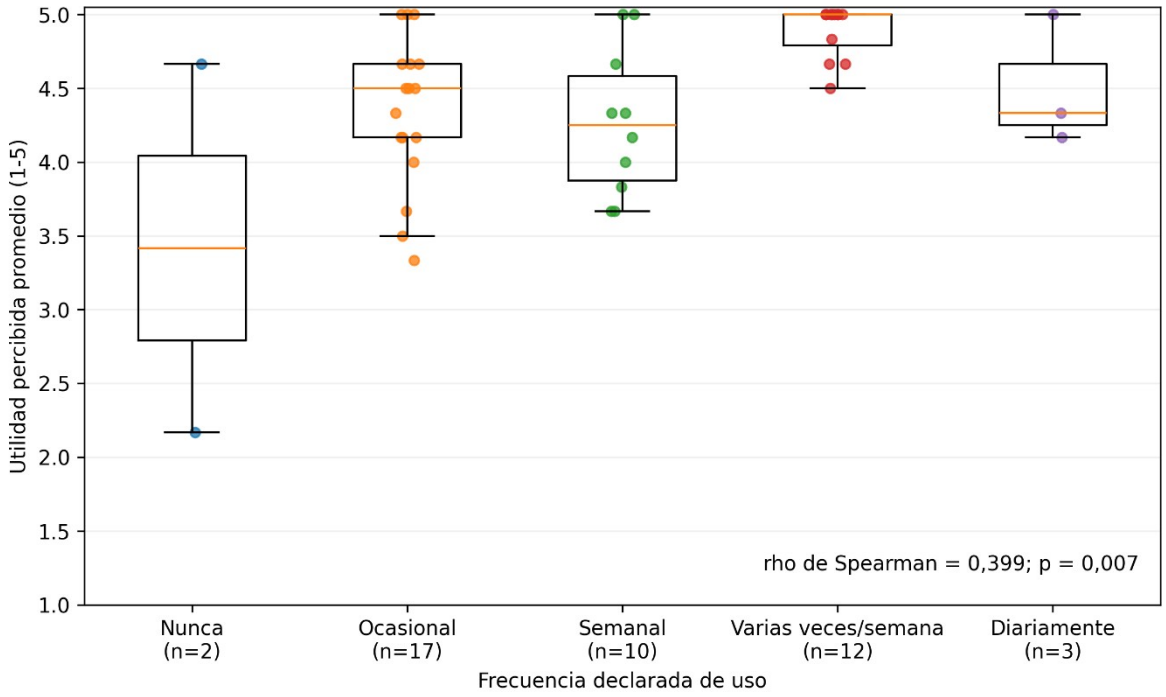


Figura 3. Distribución de la utilidad percibida según frecuencia general de uso autorreportada de tutores virtuales.

Table 4. Thematic synthesis of open responses.

Issue	Interpretation	Frequency	Brief examples
Clarity, structure, and simple explanation	The tutor is valued as a support for organizing and understanding content.	High	"simple and detailed explanations"; "order and structure"; "explains in detail".
Resolution of doubts and accuracy	Students value quick, specific answers that are aligned with the subject matter.	High	"answer my doubts"; "precise answer"; "related to the subject."
Reflection, reasoning, and error analysis	Participants describe uses or assessments linked to questioning, error review, and reasoning, with a greater presence in the 8th semester cohort.	Moderate-high	"question my learning"; "question my mistakes"; "helps me reason."
Support for study and evaluation	The tutor is used for exam preparation, summaries, flashcards, or sample tests.	Moderate	"Pre-test study"; "Test samples"; "Summaries, flashcards".
Improvements in customization, format, and visual resources	The requested improvements focus on the interface, images, a custom app, and a personalized database.	Moderate	"more personalized"; "use of images"; "more user-friendly format"; "standalone app".
Greater curricular/bibliographical adjustment	They request greater adherence to the syllabus, classes, bibliography, and teaching criteria.	Moderate	"adhering to the syllabus"; "accurate bibliographies"; "more specific in kinesiology".

The categories represent a thematic synthesis of open-ended responses and are presented to contextualize the quantitative results. References to reflection, reasoning, and clinical application are participants' own statements and not objectively evaluated educational outcomes.

3.4 Open-ended responses

Open-ended responses allowed for contextualizing the assessments expressed in the quantitative items. The most frequently mentioned aspects were grouped into clarity, structure, speed, resolution of doubts, and study support. In the 8th-semester cohort, expressions related to reflection, error analysis, reasoning, and clinical application appeared more frequently, such as the possibility of questioning one's own learning, reviewing errors, or applying knowledge in clinical situations. These expressions represent perceptions and uses described by the participants; they do not constitute direct measures of clinical reasoning, effective reflection, or transfer of learning. The suggested improvements focused on greater personalization, incorporation of visual resources, the creation of quiz-type activities, a more user-friendly format, and greater alignment with the syllabus, bibliography, and class content . Overall, both cohorts maintained a high rating for the structure and explanation, along with different descriptive patterns in some of the reported tutorial functions. The eighth-semester cohort showed higher scores in reflective questioning and justification, while the sixth-semester cohort scored higher in summarizing and outlining. However, none of the eight secondary comparisons retained statistical significance after Holm's adjustment. Therefore, these results should be interpreted as perceived functional orientations and not as confirmatory differences between cohorts or as evidence of objective changes in reflection, autonomy, or clinical reasoning.

4. Discussion

This study compared the reported preference for online tutoring styles between two cohorts of Kinesiology students and integrated indicators of perceived usefulness, study strategies, and functional orientation of use. The main finding was a marked difference between the two cohorts: in the 6th semester, the preference for the structured-explanatory tutor predominated, while in the 8th semester, the majority reported a preference for the reflective-Socratic tutor. This difference persisted after adjusting for overall frequency of use and prior experience with online tutors. However, because each cohort corresponded to a different subject and used tutors contextualized to their own curricular content and demands, the finding does not allow for the identification of an effect independent of the academic level.

Students in both cohorts consistently rated the structure, detailed explanations, and academic guidance highly. The eighth-semester cohort scored higher on items related to reflective questions and justification, while the sixth-semester cohort scored higher on summarizing and outlining. However, none of the eight secondary comparisons remained statistically significant after Holm's test. The largest difference was observed in summarizing and outlining ($\Delta = -0.44$; adjusted $p = 0.057$), suggesting it should be considered an exploratory signal rather than a confirmatory finding. These patterns may be related to differing needs for tutoring, as well as the nature of the courses, the complexity of the assignments, the assessment requirements, and the specific characteristics of the tutors. These results can be considered in relation to the literature on educational conversational agents, which has described their potential to provide immediate support, facilitate access to information, personalize interactions, and support study strategies (1, 2). At the same time, recent evidence indicates that the educational value of these systems depends on their pedagogical design and their capacity to foster self-regulation, metacognitive monitoring, and meaningful feedback (5, 6). From this perspective, virtual tutors need to be analyzed according to the pedagogical functions they aim to promote during interaction, in addition to their technological availability and student acceptance.

The greater preference expressed for the reflective-Socratic tutor in the 8th-semester cohort may be conceptually related to studies on Socratic conversational agents based on language models. In those studies, this type of mediation has been evaluated in relation to reflective thinking, critical reasoning, and academic performance under specific design conditions (3, 4). The present study did not evaluate those results, and therefore, the overlap is limited to the reported interest in features such as guiding questions, assumption testing, and answer justification. Similarly, the evidence on feedback generated by ChatGPT in clinical reasoning tasks provides a relevant framework for future evaluations, but does not allow us to infer that the preferences observed in this sample translate into improvements in reasoning or learning (13). Along the same lines, Corral-Gudino and Marcos propose the use of Socratic tutors as a strategy to preserve the student's active role and reduce the delegation of cognitive tasks to activities mediated by generative AI (14). This proposal is conceptually consistent with the reflective-Socratic approach used in the present study, although its effects on learning and autonomy were not evaluated in this sample.

In health sciences programs, the functions related to explanation, comparison, and justification are relevant because professional training progressively incorporates tasks involving information integration and the justification of clinical or therapeutic decisions. In this study, the 8th-semester cohort expressed a greater declared preference for the reflective-Socratic tutor and presented descriptively higher scores on items related to reflective questioning and justification. However, these secondary differences did not retain statistical significance after adjustment for multiple comparisons. Therefore, they can be interpreted as perceived expectations of tutorial mediation, but not as confirmatory evidence of a differentiated functional orientation or as a demonstration of greater clinical reasoning, reflection, or autonomy.

The study's contribution lies in the articulation between curricular context, tutor's pedagogical style, and higher health education. A significant proportion of the evidence on educational artificial intelligence comes from general higher education contexts, programming, or academic tasks with less clinical specificity. Kinesiology training, on the other hand, requires integrating biomedical knowledge, functional reasoning, contextual analysis, and the justification of decisions related to health problems. Considering that the evidence on AI in health professions education still presents methodological limitations and heterogeneous results (11), this work provides an initial description of how two curricular cohorts may differ in the pedagogical functions they report valuing in a virtual tutor. A scoping review published in this journal showed that the evidence on AI in clinical training is mainly concentrated in simulated or structured scenarios, presents high heterogeneity in interventions and results, and has little longitudinal follow-up (15). This situation reinforces the need to interpret the results of the present study within the specific characteristics of the subjects and the implementation analyzed.

The positive association between overall self-reported frequency of use and perceived usefulness should be interpreted with caution. The cross-sectional design prevents establishing the direction of this relationship: those who perceived greater usefulness may have used the tutors more frequently, while more frequent exposure may have increased familiarity with and appreciation of the resource. Furthermore, frequency was recorded globally and did not allow for distinguishing between the tutor used, the duration of interactions, or the relative exposure to each style. Therefore, this variable does not control for differential familiarity, nor does it allow us to establish that the reported preference arose from an equivalent comparison between the two tutors. Perceived usefulness and frequency of use also do not constitute direct evidence of deep learning, clinical reasoning, or academic performance. Recent studies have shown that certain AI supports can improve performance during a task or reduce frustration without producing equivalent learning gains, especially when the system provides final answers or displaces the student's core cognitive processes (7, 8). Likewise, over-reliance on conversational AI systems has been described as a risk to skills such as critical thinking, autonomous judgment, and decision-making (9).

The integration of quantitative results and open-ended responses allows for contextualizing the observed patterns. Exploratory composite scores showed high perceived utility, perceived performance support, and use of tutor-assisted study strategies in both cohorts. Functional indicators allowed for describing nuances in the self-reported assessments, although none of the secondary differences retained statistical significance after adjusting for multiplicity. These results correspond to self-reported contextual indicators and not to validated measurements of educational constructs. Furthermore, mentions of reflection, error analysis, reasoning, and clinical application in the open-ended responses describe the participants' perceived experience but do not verify that these processes actually occurred or improved as a result of interaction with tutors. The open-ended responses showed that students valued clarity, structure, speed, addressing doubts, and study support; in the 8th-semester cohort, references to reflection, error analysis, reasoning, and clinical application appeared more frequently. This qualitative information contextualizes the descriptive patterns, but does not confirm functional differences between cohorts nor compensate for the absence of objective educational outcomes.

The observed patterns allow us to formulate hypotheses for the future design and evaluation of online tutors. In curricular contexts where organizational and synthesis skills are valued, a greater presence of structured explanations, guided examples, and comprehension checks could be experimentally examined. In tasks requiring integration and justification, mediation based on guiding questions, comparison of alternatives, identification of assumptions, and error analysis could be evaluated. These possibilities are derived from preferences and uses reported in two specific cohorts and do not constitute recommendations on the most effective configuration or evidence of optimal tutoring progression .

Another line of research involves evaluating whether explicit guidelines on the academic use of online tutors modify how students interact with these tools. Requesting explanations, formulating guiding questions, comparing answers, and verifying information represent different interaction strategies, the educational value of which was not examined in this study. Their incorporation and evaluation can be based on the literature on self-regulation, cognitive delegation, and the responsible use of AI. In health sciences programs, this question is of particular interest due to the professional relevance of verifying information and justifying decisions, although its effects will need to be evaluated through specific research designs .

The main interpretive limitation of the study lies in the close relationship between curriculum cohort, educational level, subject, and the specific tutor configuration. Although both cohorts had one structured-explanatory tutor and one reflective-Socratic tutor, each pair was constructed from the curriculum corpus specific to the corresponding subject. Consequently, the observed association between cohort and tutoring style preference cannot be attributed solely to the student's educational stage. It may also be influenced by the complexity of the tasks, the evaluative demands, the type of content, familiarity with the tool, and the specific implementation characteristics of each tutor. The cross-sectional design prevents establishing causality or examining intraindividual changes throughout the academic trajectory.

All variables analyzed consisted of self-reported preferences, usefulness, study strategies, and functional orientation of use. Objective measures of learning, academic performance, quality of justification, clinical reasoning, reflective thinking, cognitive autonomy, and transfer to authentic tasks were not included. Consequently, participants' references to reflection, error analysis, or clinical application do not allow us to conclude that these processes occurred, improved, or were caused by the online tutor. Nor can we establish the educational effectiveness or superiority of one tutoring style over another.

The questionnaire was developed specifically for this implementation and was not subjected to a formal content, construct, or criterion validation process. The primary variable consisted of a closed-ended question with stated preference, while secondary outcomes were constructed using predefined item groupings. Internal consistency analyses provide descriptive information about the behavior of these scores in the studied sample, but they do not demonstrate unidimensionality nor do they replace a full psychometric evaluation. Perceived usefulness showed $\alpha = 0.790$, with corrected item-total correlations between 0.401 and 0.694. The study strategies score showed more modest internal consistency, with $\alpha = 0.683$ and corrected item-total correlations between 0.282 and 0.658. The removal of any item resulted in coefficients between 0.562 and 0.690, indicating that no single item explained the lower observed homogeneity. This characteristic may be related to the diversity of strategies included and reinforces its interpretation as a contextual exploratory grouping. Perceived performance support was comprised of two items, presented $\alpha = 0.867$ and an item correlation of $r = 0.770$, and should therefore be interpreted as a brief composite score. Consequently, these results express contextual indicators of perception and reported use, not standardized measurements of established educational constructs .

Another limitation relates to the size and characteristics of the sample. The study used a small, non-probabilistic sample from a single program and institution, which restricts the generalizability of the findings. Although a different gender distribution was observed between cohorts, this variable was not included in the main model to maintain parsimony in the analysis with a small sample size. Its potential influence on stated preferences should be evaluated in studies with larger sample sizes.

The voluntary and naturalistic use of tutoring constitutes another significant limitation. Although students had access to both tutors, direct comparison was not required, nor were the

timing, sequence, or number of interactions standardized. Objective usage data were also not recorded to identify the tutor used, the duration of each session, or the cumulative exposure to each style. Self-reported frequency was generalized and, therefore, did not allow for controlling for differential familiarity or specific intensity of use. Under these conditions, the stated preference may be influenced by self-selection, particular academic needs, prior motivation, or unequal exposure.

Secondary analyses included eight comparisons between cohorts, for which Holm's sequential procedure was applied. After controlling for multiplicity, no difference remained statistically significant. Although the item for preparing summaries or outlines showed the largest difference, its adjusted p-value was 0.057. These results require caution due to the sample size, the self-reported nature of the variables, and the exploratory nature of the analysis.

Future research should incorporate crossover or experimental designs in which the same participants use both tutoring styles with equivalent content, counterbalanced sequences, and comparable exposure times. It would also be necessary to record, with explicit consent and ethical safeguards, the number of sessions, the duration of interactions, the tutor used, and the dialogue patterns. These measures would allow researchers to distinguish informed preference from differential familiarity and examine the relationship between specific forms of interaction and objective learning outcomes, simulated clinical reasoning, or the quality of justification. It would also be pertinent to develop and evaluate instruments with formal evidence of validity and consistency to measure the educational dimensions of interest.

In summary, the study provides contextualized evidence on the stated preferences and valued tutorial functions in two Kinesiology curriculum cohorts. Both cohorts showed a high valuation of structure, guidance, and explanation, along with a marked difference in preference for tutorial style. Secondary functional indicators showed some descriptive patterns, but no comparison retained statistical significance after adjusting for multiplicity. The findings allow us to formulate questions and hypotheses about the pedagogical configuration of conversational tutors, but they do not support recommendations regarding their effectiveness nor demonstrate effects on learning, reflection, professional autonomy, or clinical reasoning .

5. Conclusions

- In the two curricular cohorts studied, marked differences were observed in the reported preference for virtual tutoring styles. In the 8th semester cohort, the preference for the reflective-Socratic tutor predominated, while in the 6th semester cohort, the preference for the structured-explanatory tutor predominated.
- Structure, detailed explanation, and academic guidance were valued in both cohorts. Different descriptive patterns were observed in reflective questions, justification, and the creation of summaries or outlines, but none of the eight secondary comparisons retained statistical significance after Holm's adjustment. These results describe self-reported functional preferences and orientations; the study did not assess objective outcomes of learning, clinical reasoning, reflection, autonomy, or educational transfer .
- Since each cohort corresponded to a different subject and used tutors contextualized to specific curricular demands, the observed differences cannot be attributed solely to the educational level. Furthermore, exposure to each tutor was voluntary, non-standardized, and recorded only through a general self-reported frequency; therefore, it cannot be assumed that the stated preference resulted from an equivalent comparison between the two styles.
- The observed patterns suggest that the relationship between tutorial configuration, curricular context, task characteristics, and expected level of autonomy warrants specific evaluation. Future cross-sectional, longitudinal, or experimental studies, with equivalent exposure and

objective measures of learning and performance, should determine whether any configuration produces different educational outcomes .

Funding : There was no funding.

Declaration of conflict of interest : The author declares that he has no conflict of interest.

Use of artificial intelligence in the research process : ChatGPT Plus was used to configure personalized virtual tutors using the "Create GPT" function, restricted to the learning outcomes, content, activities, and official bibliography of the participating courses. Students had free access to a structured-explanatory tutor and a reflective-Socratic tutor, without any conversations held with these tools being collected or analyzed. In the analytical phase, Codex and ChatGPT were used solely for technical support in generating and debugging Python code for data processing and statistical calculations. No student identification data was entered into the artificial intelligence tools. The methodological decisions, selection of results, statistical interpretation, final drafting, and conclusions were made and verified by the author, who assumes full responsibility for the content of the manuscript.

Author's contributions : ICM carried out the conceptualization, methodology, design of the virtual tutors, development of the data collection instrument, ethical management, data collection, formal analysis, interpretation of results, writing of the original draft, revision and editing of the manuscript, visualization of results and approval of the final version.

Annexes: Supplementary material 1 (Pedagogical architecture and core instructions for virtual tutors) and 2 (Data collection instrument).

6. References

1. Okonkwo CW, Ade-Ibijola A. Chatbots applications in education: a systematic review. *Comput Educ Artif Intell* . **2021** , 2, 100033. https://doi.org/10.1016/j.caeai.2021.100033 .
2. Abdallah N, Katmah R, Khalaf K, Jelinek HF. Systematic review of ChatGPT in higher education: navigating impact on learning, wellbeing, and collaboration. *Soc Sci Humanit Open* . **2025** , 12, 101866. https://doi.org/10.1016/j.ssaho.2025.101866 .
3. Xi L, Zhang Y, Wang Q. Investigating the effects of an LLM-based Socratic conversational agent on students' academic performance and reflective thinking in higher education. *Comput Educ* . **2026** , 241, 105494. https://doi.org/10.1016/j.compedu.2025.105494 .
4. Fakour H, Imani M. Socratic wisdom in the age of AI: a comparative study of ChatGPT and human tutors in enhancing critical thinking skills. *Front Educ* . **2025** , 10, 1528603. https://doi.org/10.3389/feduc.2025.1528603 .
5. Chang DH, Lin MP-C, Hajian S, Wang QQ. Educational design principles of using AI chatbot that supports self-regulated learning in education: goal setting, feedback, and personalization. *Sustainability* . **2023** , 15, 12921. https://doi.org/10.3390/su151712921 .
6. Guan R, Raković M, Chen G, Gašević D. How educational chatbots support self-regulated learning? A systematic review of the literature. *Educ Inf Technol* . **2025** , 30, 4493 – 4518. https://doi.org/10.1007/s10639-024-12881-y .
7. Bassner P, Lenk-Ostendorf B, Beinstingel R, Wasner T, Krusche S. Less stress, better scores, same learning: the dissociation of performance and learning in AI-supported programming education. *Comput Educ Artif Intell* . **2026** , 10, 100537. https://doi.org/10.1016/j.caeai.2025.100537 .
8. Bastani H, Bastani O, Sungu A, Ge H, Kabakcı Ö, Mariman R. Generative AI without guardrails can harm learning: evidence from high school mathematics. *Proc Natl Acad Sci USA* . **2025** , 122, e2422633122. https://doi.org/10.1073/pnas.2422633122 .
9. Zhai C, Wibowo S, Li LD. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review . *Smart Learn Environment* . **2024** , 11, 28. https://doi.org/10.1186/s40561-024-00316-7 .
10. Campos Montenegro I, Martínez Mena M. AI literacy and subjective experience with a GPT-based virtual tutor in Kinesiology students: a cross-sectional study. *Rev Esp Edu Med* . **2026** , 7(4). https://doi.org/10.6018/edumed.715291 .
11. Feigerlova E, Hani H, Hothersall-Davies E. A systematic review of the impact of artificial intelligence on educational outcomes in health professions education. *BMC Med Educ* . **2025** , 25, 129. https://doi.org/10.1186/s12909-025-06719-5 .

12. Izquierdo-Condoy JS, Arias-Intriago M, Tello-De-la-Torre A, Busch F, Ortiz-Prado E. Generative artificial intelligence in medical education: enhancing critical thinking or undermining cognitive autonomy? *J Med Internet Res* . **2025** , 27, e76340. https://doi.org/10.2196/76340 .
13. Çiçek FE, Ülker M, Özer M, Kıyak YS. ChatGPT versus expert feedback on clinical reasoning questions and their effect on learning: a randomized controlled trial. *Postgrad Med* J. **2025** , 101, 458 – 463. https://doi.org/10.1093/postmj/qgae170 .
14. Corral-Gudino L, Marcos M. Writing in medical education in the age of generative artificial intelligence: recommendations to foster creativity and critical thinking in a generative AI-assisted learning environment. *Rev Esp Edu Med* . **2026** , 7(3). https://doi.org/10.6018/edumed.708901 .
15. Bonilla Mejia JJ, Cortés Fuenzalida TD, Polanco Aliaga DH, Martínez Carrillo C, Herrera Alcaíno AA. Artificial intelligence in the practical clinical training of undergraduate medical students: a review of the scope of applications, results and gaps. *Rev Esp Edu Med* . **2026** , 7(3). https://doi.org/10.6018/edumed.710071 .



© 2026 University of Murcia. Submitted for open access publication under the terms and conditions of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 Spain License (CC BY-NC-ND) (http://creativecommons.org/licenses/by/4.0/)