SAMYRAD 2026

SMART ALGORITHMS & MODELS: YOUR REACH TO ADVANCED DIALOGUES · 5–6 OCTOBER 2026 · SEVILLE, SPAIN
SAMYRAD OPEN-ACCESS PROCEEDINGS | PUBLISHER: INTELLIGENT NUDGE

PAPER • OPEN ACCESS

Comparing AI Interaction Logs and Classroom Analytics for Detecting Conceptual Difficulties in Statistics

Fulya Kula
(e-mail: f.kula@utwente.nl)
Applied Mathematics Department, Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS), University of Twente, Enschede, The Netherlands

Article DOI: https://doi.org/10.68286/samyrad2026.370

ABSTRACT
Generative AI tools are increasingly used by students of statistics, but it is unclear whether the questions students ask them reveal the conceptual difficulties that classroom assessment already detects. Classroom polling surfaces a misconception only when a question happens to probe it, so a source available throughout a course could in principle complement it. This exploratory case study asks which topics generate the most AI interactions, which show the highest classroom-response error rates, and how far the two sources agree. We compared 728 AI-log items (162 conversations, 58 students) from a curriculum-bounded platform with 18 Wooclap items from three tutorials in a second-year statistics course at a Dutch university, with final-exam results as a third reference point, using a shared six-topic coding scheme. AI-log volume did not align with Wooclap error rates: among the three topics present in both sources, Type I/II errors and statistical power had the fewest AI conversations (4) but the highest error rate (57.4%), and point estimation, with the highest exam error rate among comparable items, had no Wooclap item. The comparison is descriptive and limited to one course. It suggests that AI-log volume shows where attention concentrates rather than where students struggle, and that the two sources are best read together.

KEYWORDS generative AI, statistics education, interaction logs, classroom response systems, conceptual difficulties

How to cite: Kula, F. (2026). Comparing AI interaction logs and classroom analytics for detecting conceptual difficulties in statistics. SAMYRAD 2026 Open-Access Proceedings. Intelligent Nudge. https://doi.org/10.68286/samyrad2026.370

Introduction

Generative AI tools are increasingly available to students studying statistics, but whether the questions students choose to ask such a tool reveal anything about their conceptual understanding, rather than simply serving as a convenient homework aid, is not yet well established. Statistics education research has documented a range of persistent misconceptions in how students reason about inference [1]. One well-studied example concerns confidence intervals: Hoekstra et al. [2] show that misinterpreting a confidence interval as a probability statement about the population parameter persists even among trained researchers, not only introductory students. Findings of this kind give a new data source something concrete to test itself against: if the topics students raise with an AI tool point toward misconceptions already documented in this literature, that convergence is reassuring; if the two diverge, the divergence itself says something about what each method can and cannot detect.

The three tutorial sessions examined in this study were structured around think-pair-share, a cooperative-learning

technique in which students first consider a question individually, then discuss it with a partner, before a wider class discussion [3], [4]. The technique has been applied specifically in statistics instruction, where it forms part of a broader set of active-learning methods shown to improve engagement over traditional lecturing [5]. More broadly, across disciplines, recent work has also questioned how essential the final "share" step is to the technique's benefit, and has called for research into alternative ways of closing a think-pair cycle [6]. The present course's tutorials adapted the format accordingly: alongside discussion with a partner and the class, students could also consult the AI platform directly, a variant we refer to as think-pair-and-AI-share. Within the same sessions, Wooclap, an interactive classroom response system, was used to poll students on the same material: students answer a question on their own device after scanning a QR code, and the instructor then displays the class-wide response distribution and the correct answer, which anchors discussion during the share phase of each tutorial. This built on a longer tradition in which structured peer discussion measurably improves performance on in-class concept questions [7], and on more recent evidence that Wooclap specifically supports engagement and achievement in comparable higher-education settings [8], [9].

What such classroom-based methods cannot easily capture is a misconception a student never has the opportunity to reveal, because no polling question that particular session happened to probe it. An AI tool that a student can query at any point, on any topic, in principle fills exactly that gap. Whether it does so in practice, generating conceptually meaningful questions rather than off-task use, is less established. A semester-long study of a physics course chatbot showed that recovering interpretable patterns of student reasoning from raw conversational logs is possible, but that doing so requires deliberate methodological work [10]. The AI log examined here spans a comparably extended period, and this study faces the same underlying challenge: showing that the signal recovered from student-authored prompts reflects genuine conceptual engagement, not an artifact of how the data happened to be logged.

This paper compares AI interaction logs and Wooclap classroom response data collected within the same second-year undergraduate statistics course, both generated in parallel within the same three think-pair-and-AI-share tutorial sessions, Tutorials I, II, and III. It addresses three questions: which statistical topics generate the most AI interactions, which topics show the highest rates of incorrect Wooclap responses, and to what extent the two sources agree on where student difficulty concentrates. Final exam records, including two items whose scenarios were numerically identical to specific Wooclap questions, provide a third, independent comparison point. Reaching this comparison required substantial preliminary data-cleaning work, documented in the Methodology: the raw AI-log export required structural reconstruction before it could be parsed at all, contained interactions from a non-student test account, and included conversations scoped to courses other than the one under study. These steps are reported in detail because similar problems are likely to come up in future studies using AI interaction logs. The paper first reviews what is known about documented misconceptions in statistical inference, classroom response systems, and AI interaction logs and generative AI tools in statistics education, before describing the course context and data sources, reporting results for each of the three questions above, and discussing what the alignment and divergence between data sources suggests, together with the limitations of a design in which AI-tool use was voluntary rather than assigned.

Literature review

Documented Misconceptions in Statistics Education

Statistics education research has documented a range of persistent misconceptions in how students reason about data and inference [1]. One well-studied example concerns confidence intervals: Hoekstra et al. [2] show that misinterpreting a confidence interval as a probability statement about the population parameter persists even among trained researchers, not only introductory students. A second example concerns point estimation and sampling-distribution reasoning. A review of research on statistical sampling instruction finds that students, with or without instructional simulations, often fail to grasp that a sampling distribution describes the spread of a statistic across many samples, not a single sample, a distinction made harder because both are commonly visualized the same way, as a histogram [11]. A separate study notes that misconceptions research of this kind has relied on a fairly narrow set of methods, mainly concept inventories and a smaller number of think-aloud interview studies, and argues that other methods for surfacing student reasoning remain underused [12]. This gap is directly relevant here: an AI interaction log is one such underused method, and this study examines whether it can surface the same kinds of documented difficulty.

Classroom Response Systems and Think-Pair-Share in Statistics Instruction

The tutorials examined in this study were built around think-pair-share, a cooperative-learning technique introduced by Lyman [3], [4], in which students consider a question individually, discuss it with a partner, and then take part in wider class discussion. The technique has been applied specifically in statistics instruction as part of a broader set of active-learning methods shown to improve engagement over traditional lecturing [5]. More recent work has questioned how essential the final "share" step is to the technique's benefit and has called for research into alternatives [6]. Smith et al. [7] provide foundational evidence that structured peer discussion, paired with instructor explanation, measurably improves performance on in-class concept questions, establishing the pedagogical logic behind polling-based tutorials of the kind used here. Wooclap specifically has been shown to support engagement and achievement in comparable higher-education settings, including economics [8] and chemical engineering [9] courses. This literature establishes that classroom response data reliably surfaces topic-level error rates in real time, but it does not typically compare those error rates against an independent, self-directed data source such as AI tool use, which is the comparison this study attempts.

AI Interaction Logs and Generative AI Tools in Statistics Education

Whether AI interaction logs reveal genuine conceptual engagement, rather than mostly off-task or administrative use, is not yet well established. A semester-long study of a physics course chatbot showed that recovering interpretable patterns of student reasoning from raw conversational logs is possible but requires deliberate methodological work, in that case an unsupervised clustering pipeline [10]. Within statistics specifically, GenAI's role has so far been discussed mainly through student self-report and course-design studies rather than through analysis of behavioral log data. A survey-based study spanning a full semester found that students rate ChatGPT favorably across several dimensions of academic support, with writing assistance and exam preparation emerging as the strongest predictors of perceived overall

effectiveness [13]. A study of project-based statistics courses found that students report favorable views of course projects generally, and that many had independently incorporated ChatGPT into their coursework, describing perceived benefits, though this use was student-initiated rather than part of the course design [14].

Ellis and Slade [15] outline considerations for using ChatGPT to enhance statistics and data science instruction. Working with actual student writing samples rather than self-report, one study analyzes a corpus of over 1,600 undergraduate data-analysis reports from 2021 to 2025 and finds that students' writing style has shifted measurably since large language models became widely available, becoming both more similar to LLM-generated text and, at the same time, more similar to expert statistical writing [16]. A related study shows that LLM-generated and human-generated statistical writing nonetheless remain stylistically distinguishable, informing how instructors might guide LLM use in writing-based assessments [17]. None of this work, to our knowledge, has used the content of AI interaction logs themselves as a diagnostic signal for topic-level conceptual difficulty in a statistics course, cross-checked against an independent classroom data source. That is the gap this study addresses.

Methodology

Context and Participants

This study was conducted in a second-year undergraduate statistics course at a university in the Netherlands. Around 400 students were registered, and 265 completed the final exam. The course used two tools together: Alexandria [18] (since renamed Odea), a curriculum-bounded generative AI platform restricted to course materials, and Wooclap.

Three tutorials ran as think-pair-and-AI-share sessions, combining both tools. Tutorial I came after descriptive statistics had been introduced and covered point estimation and confidence intervals, with 45 Wooclap participants. Building on this, Tutorial II covered basic concepts of hypothesis testing, with 27 participants. Tutorial III followed and covered two sample hypothesis tests, with 16 participants. Later course content, including normality assessment, categorical data analysis, and linear regression, came after these three tutorials and was not part of any think-pair-and-AI-share session.

Ethical approval was granted by the university's faculty ethics committee. Students opted in at first login to Alexandria and at the start of each Wooclap session.

Data Sources

AI interaction logs.

Alexandria exported timestamped question-answer pairs, item type (free-text query, or a request to explain highlighted text or an image), conversation grouping, and course scope. Three cleaning steps came before analysis. One account showed a usage pattern inconsistent with genuine student use: the same prompt repeated 138 times, starting before the course's first lecture. This account was removed. The exports also included interactions from at least three other modules, so records were restricted to those scoped to this course. A further six conversations referenced a different course or module by name in their actual content, not just their scope tag, and were excluded too. The resulting dataset held 728 interaction items across 162 conversations from 58 students, spanning the full course period.

Wooclap responses.

These data came from the same three think-pair-and-AI-share tutorials described above, Tutorials I, II, and III, producing 18 usable items once one consent or demographic item per session was set aside: 7 multiple-choice items on confidence intervals (Tutorial I), 8 on hypothesis-testing fundamentals, Type I and II errors, and statistical power (Tutorial II), and 3 open-response items on two-sample hypothesis testing (Tutorial III). Response counts per item ranged from 12 to 45, depending on attendance. The three open-response items were not machine-graded and are treated separately from the rest. Table 1 shows two representative items, one from Tutorial I and one from Tutorial III, to illustrate the item format; these two also reappear later as the items whose scenarios were reused verbatim on the final exam.

Exam records.

Final exam grades for all 265 students covered 11 questions and 43 points total. The exam was written and closed book: no electronic devices or internet access were permitted, and students were limited to a simple, non-graphing calculator. The exam had four versions with identical content and point structure; only the ordering of multiple-choice options and the wording of distractors differed, so scores are directly comparable across versions. Two of the eleven questions matched Wooclap content exactly, reusing the same scenario, one on confidence intervals from Tutorial I, and one on two-sample procedures from Tutorial III. These two gave an exact-match comparison point in addition to the broader topic-level comparison.

Table 1. Example WOOCLAP items

Session	Topic	Item text
Tutorial I	Confidence intervals	A tech company is developing a new wearable device. The device's older model had a well-documented population standard deviation σ = 2.0 hours for battery lifetime. However, after updates to the hardware and firmware, engineers suspect this value may no longer apply. A tester measures the battery lifetime of 18 prototypes and obtains: mean = 204.3 hours, sd = 2.1 hours. Which of the following correctly expresses the 95% confidence interval for the true mean battery lifetime?
Tutorial III	Two-sample procedures	A cybersecurity team is analyzing phishing-click rates in two departments of a company: Department A: 3 of 40 employees clicked the phishing link. Department B: 12 of 200 employees clicked the phishing link. You want to test whether the phishing-click rate in Department A differs from that in Department B, at significance level α = 0.05.

Topic Coding

All three sources were coded against a shared set of topics, built from what the three tutorials actually covered: descriptive statistics, point estimation, confidence intervals, hypothesis-testing fundamentals, Type I and II errors and statistical power, and two-sample procedures. Later syllabus content was left out of the coding scheme and out of cross-source comparison, since neither tool could have shaped performance on material students had not yet seen.

AI-log items were coded at the conversation level, not the individual turn. A conversation's topic came from the first turn

with a topic-specific keyword match, and every later turn in that conversation inherited the same code. Many follow-up turns, things like "help me answer part a," carry no topical keyword of their own but still belong to the conversation that opened with one. Coding turn by turn left about half of all items with no topic match; coding conversation by conversation cut that to under a fifth. At the conversation level, however, 77 of the 162 conversations (47.5%) matched no topic keyword at any point and were coded "OTHER". A manual read suggests most of these were not conceptual (see Results), but some may have used vocabulary the keyword scheme did not capture. Keyword lists were built from the vocabulary specific to each of the six tutorial topics and are not included here in full; the approach can be shared on request.

Wooclap items were coded individually, since each one was a single self-contained scenario. Exam items were mapped to topics by reading their content directly rather than by keyword matching, since there were only 11 of them and several combined more than one topic. Three questions mixed topics under one score and were dropped from topic-level comparison: one combining nine true/false statements across descriptive statistics, confidence intervals, and Type I/II errors; one combining confidence-interval construction with hypothesis formulation and a statistical power calculation; and one combining a variance-comparison test with a sample-size calculation. Three further questions covered later course content (normality assessment, categorical data analysis, and linear regression) and fell outside the coding scheme. The remaining five questions each addressed a single topic, point estimation, hypothesis testing, two-sample comparison (twice, across two different questions), and confidence intervals, and were kept for the comparison. All topic coding, the keyword matching applied to the AI log and the manual mapping of exam questions to topics, was carried out by the author; no second coder independently verified these assignments.

Analytic Procedure

Three questions were addressed descriptively. Which topics generate the most AI interactions was answered by ranking topics on conversation count and item count, with conversation count treated as primary, since item counts are sensitive to a handful of long conversations (median 2 items per conversation, maximum 57). Which topics show the highest Wooclap error rates was answered by aggregating respondent-level correctness within each topic across the 15 graded multiple-choice items; the 3 open-response items were read qualitatively instead. How far the two sources align was addressed descriptively too: topics were ranked by AI-log conversation volume next to their Wooclap error rate and, where a matching single-topic exam item existed, their exam error rate. No correlation coefficient was computed, given only six topic categories to work with. The two exam items with an identical Wooclap scenario were treated as a separate, higher-confidence comparison rather than folded into the broader topic-level one.

Limitations of the Data

Three things about the data are worth flagging before the Results. First, the AI log covers the whole course, not just the three tutorials; only 89 of 728 items fall within the three tutorial moments. Most AI-log activity is independent study outside these sessions, not in-session use directly comparable to a Wooclap snapshot. Second, course-scope tags in Alexandria are not a perfect guide to content, since a few tagged conversations turned out to reference other coursework by name. The exclusion described above catches the clear cases, but subtler off-topic content without an explicit course name could still remain. Third, because a conversation's topic is set by its first keyword match and carried forward, a conversation that opens on one topic and later shifts to another is coded entirely under the opening topic, so topic drift within a conversation is not captured.

Results

Which statistical topics generate the most AI interactions?

Table 2 breaks down the 728 retained AI-log items (162 conversations, 58 students) by topic; the percentage column expresses each topic's share of the 85 conversations that received a topic code, excluding the 77 conversations coded "OTHER".

Table 2. AI-log volume by topic

Topic	Conversations	Items	% of coded conversations
Descriptive statistics	33	231	38.8%
Point estimation	18	122	21.2%
Confidence intervals	15	79	17.6%
Hypothesis-testing fundamentals	12	121	14.1%
Type I/II errors and statistical power	4	35	4.7%
Two-sample procedures	3	12	3.5%
"OTHER"	77	128	—

Descriptive statistics drew the most AI-log activity, followed by point estimation, confidence intervals, and hypothesis-testing fundamentals. Type I/II errors and statistical power and two-sample procedures trailed well behind. A large share of conversations, 77 of 162, matched no topic keyword and were coded "OTHER". A manual read of this category showed it was predominantly course logistics, platform-navigation questions, and requests to explain a highlighted passage or image; a smaller share could not be ruled out as conceptual content using vocabulary the keyword scheme did not capture, though a precise breakdown was not retained. Item counts per conversation varied a lot too, from 1 to 57, with a median of 2. This spread means item counts overweight a handful of long conversations, so the conversation counts in Table II are the more reliable guide to where attention concentrated. Descriptive statistics and point estimation, the two earliest topics on the syllabus, together made up 60% of coded activity. Students seem to have leaned on the tool more heavily early in the course.

Which topics show the highest rates of incorrect Wooclap responses?

Error rates were high across all three topics with graded Wooclap items. Type I/II errors and statistical power had the highest error rate, 57.4% across 61 respondent-answers on 4 items. Hypothesis-testing fundamentals followed closely at 56.1% across 66 respondent-answers on 4 items. Confidence intervals came in at 50.4% across 119 respondent-answers on 7 items, including the example item shown in Table I. The three two-sample items were open-response and not machine-graded, so they sit outside this comparison; their content comes back later, via the matching exam item. Within confidence intervals, error rates ranged widely by item, from 12.5% on a margin-of-error formula question to 66.7% on a question asking students to interpret a confidence interval as

a probability statement about the population mean. The topic-level average hides this spread: the highest error rate in the topic belonged to the item on interpreting a confidence interval as a probability statement, although a single item cannot show that this one misconception drives the topic-level rate.

How far do the two sources agree?

Before comparing the three sources directly, it is worth repeating that AI-log volume reported here accumulates across the full course period, while the Wooclap and matching exam figures reflect three specific tutorial sessions; the comparison that follows therefore sets a cumulative, full-term measure against session-specific snapshots, not two measures taken over the same window. Table 3 lines up AI-log volume, Wooclap error rate, and exam error rate by topic. The exam column uses five exam questions that each addressed a single topic cleanly; three further exam questions mixed several topics under one score and are left out here, as explained in Methodology. Descriptive statistics, despite generating the highest AI-log volume of any topic, is not shown in Table 3: unlike point estimation, it has neither a matching Wooclap item nor a single-topic exam question, so no comparison point exists for it in either source, not because it was judged less relevant.

Table 3. AI-log volume, Wooclap error rate, and exam error rate

Topic	AI conversations	Wooclap error rate (n)	Exam error rate (n)
Point estimation	18	— (no matching Wooclap item)	70.4% (n = 265)
Confidence intervals	15	50.4% (n = 119)	38.4% (n = 265), exact scenario match
Hypothesis-testing fundamentals	12	56.1% (n = 66)	25.2% (n = 265)
Type I/II errors & statistical power	4	57.4% (n = 61)	—
Two-sample procedures	3	N/A (open response)	12.5% / 40.1% (n = 265 each), exact scenario match

Two things stand out. First, point estimation had the second-highest AI-log volume of any topic and the highest exam error rate of the four topics with exam data in Table III, but no Wooclap item across the three sessions touched it at all. The topic students engaged with most in the AI tool and struggled with most on the exam was invisible to the classroom-response data, simply because none of the three sessions happened to include a question on it. Second, two-sample procedures had the lowest AI-log volume of any topic (only 3 conversations, so this comparison should be read cautiously), yet its two matching exam items told very different stories: 12.5% error on one two-sample scenario, 40.1% error on the other, the same scenario also used in a Wooclap session. Low AI-tool engagement did not mean uniformly strong performance on that topic.

For the three topics present in both the AI log and Wooclap, confidence intervals, hypothesis-testing fundamentals, Type I and II errors and statistical power, more AI-log conversations did not mean a lower Wooclap error rate. If anything the opposite held: the topic with the fewest AI-log conversations, Type I and II errors and statistical power (4 conversations, the smallest of the three), had the highest Wooclap error rate of the three; given that small base, this specific comparison should be read cautiously. AI-log volume, on its own, does not tell you where the difficulty is.

Where an exam question used the exact same scenario as a Wooclap item, the exam error rate for the confidence-interval item was lower than the corresponding Wooclap figure. On the confidence-interval item shown in Table I, the exam error rate for the identical scenario (38.4%) was lower than the confidence-interval topic average in the tutorial (50.4%), though a topic average is not directly comparable to a single item. This fits with continued exposure between the tutorial and the exam, but this study cannot separate the AI tool's specific role in that improvement from ordinary revision or other exam preparation.

Discussion

Convergence and Divergence Between the Two Sources

The Introduction posed a simple test: if the topics students raised with the AI tool pointed toward misconceptions already documented in the statistics education literature, that would be reassuring; if the two diverged, the divergence itself would be informative. This was not framed as a fourth research question, it is a lens for interpreting the three results already reported, not a separate finding to report on its own. Under that lens, the answer is divergence. For confidence intervals, hypothesis-testing fundamentals, and Type I and II errors and statistical power, the only three topics present in both datasets, the ranking by AI-log conversation count ran opposite to the ranking by Wooclap error rate. The topic with the fewest AI-log conversations, Type I and II errors and statistical power, had the highest Wooclap error rate of the three.

Several explanations seem plausible, and the data collected here cannot distinguish between them. Students who struggled most with Type I and II error and statistical power reasoning may not have recognized their own difficulty clearly enough to ask about it. It is equally possible that attendance varied across the three tutorials, or that fewer students had the AI tool open during that particular exercise, so low volume on this topic may reflect opportunity rather than confidence. Alternatively, confidence-interval content, despite a lower classroom-response error rate, may have drawn more follow-up questions because its correct interpretation is subtle enough that a single classroom session leaves residual uncertainty; Hoekstra et al. [2] found that confidence-interval misinterpretation persists even among trained researchers, which fits this third explanation. A broader asymmetry sits behind all three: Wooclap participation was compulsory and tied to the three tutorial sessions, while AI-tool use was voluntary and open across the whole course, so the two sources were never sampling the same decision to engage in the first place, itself a plausible driver of the divergence, separate from topic difficulty. What the data do show is that AI-log volume and classroom-response error rate measure different things, volume reflects inquiry frequency, error rate reflects incorrect responses, and the one cannot be treated as an unproblematic stand-in for the other, a conclusion only visible because the Wooclap comparison existed at all.

Point Estimation

Point estimation generated the second-highest AI-log conversation count of any topic, and on the exam, the highest error rate of the five single-topic questions. The relevant exam question asked students to identify the value of a scaling constant that made a sample mean an unbiased estimator, and the value that minimized mean squared error (MSE), a

common measure of estimator accuracy that combines both bias and variability. No Wooclap item across the three sessions touched point estimation at all, since Wooclap items were written session by session to match each tutorial's planned content, and none of the three happened to include one on estimator bias or MSE.

This gap matches a pattern already documented in the statistics education literature. A review of research on statistical sampling instruction finds that students, with or without instructional simulations, often fail to grasp that a sampling distribution describes the spread of a statistic across many samples, not a single sample [11]. A separate study notes that student misconceptions in this area have historically been studied through a narrow set of methods, mainly concept inventories and a smaller number of think-aloud studies, leaving open questions about how such difficulties surface in less structured settings [12]. An AI interaction log is one such less structured setting. Here, point estimation drew substantial AI-log activity but had no Wooclap item, so a Wooclap-only design would have given no classroom-response signal on it. Whether that activity reflects the estimator-reasoning difficulties described in this literature cannot be determined from volume alone.

The two exact scenario matches between the exam and the Wooclap sessions add a further point of comparison. One was the confidence-interval item shown in Table I. The other was a two-sample scenario that appeared in both a Tutorial III Wooclap item and, later, on the exam. For the confidence-interval item the exam error rate was lower than the tutorial topic average, while for the two-sample item no graded Wooclap rate exists to compare with. This is consistent with continued exposure between the tutorial and the exam, but this study's design cannot separate the AI tool's specific role in that improvement from ordinary revision, lecture reinforcement, or studying with classmates.

Practical Lessons for Future Work

The dataset built for this study needed considerably more preprocessing than the raw export initially suggested. This belongs in the Discussion rather than the Methodology because the point here is not what this study did, that is already described in the Methodology, but what the experience implies for anyone attempting similar work in future: these are exactly the problems worth checking for before treating an AI-log export as analysis-ready. The raw export required reconstruction of the underlying file structure before any of it could be read. One account behaved nothing like a student, repeating a single prompt 138 times starting before the course had begun, and needed to be removed outright. Course-scope metadata missed content from at least three other modules and, on closer reading, a handful of conversations that named a different course directly in their text despite carrying the correct scope tag. Existing chatbot-log studies describe some preprocessing, typically isolating student-authored text from bot responses, but a broken export, unreliable scope tags, and a non-student account are problems that literature has not yet described.

Limitations

A few limitations qualify what can be claimed here. The two datasets were not collected under comparable conditions: Wooclap participation was compulsory and tied to three specific sessions, while AI-tool use was voluntary and spread across the whole course, with only 89 of 728 retained items falling within the three tutorial moments. Students who chose to use the AI tool are not necessarily representative of the class, and motivation, prior ability, or study habits could plausibly explain both AI-tool use and exam performance. Individual students could not be linked by identity across the AI log, Wooclap, and exam records, so every comparison in this study is aggregate and topic level only, not traceable to a single student's path across all three sources. Topic and tutorial session are also fully confounded in this design: estimation and confidence intervals were covered only in Tutorial I, hypothesis testing and Type I/II errors only in Tutorial II, and two-sample procedures only in Tutorial III, so topic-level differences cannot be separated from differences in attendance or cohort composition across the three sessions. Sample sizes were also small: Wooclap items drew between 12 and 45 responses, and with 12 responses the 95% margin of error on an error rate can reach roughly ±28 percentage points, so many item-level differences lie within measurement noise. The topics with only 3 or 4 AI-log conversations (two-sample procedures, and Type I and II errors and statistical power) are especially fragile. The topic-coding scheme relies on keyword matching at the conversation level, which cannot distinguish surface vocabulary from real conceptual content. The keyword scheme was not validated against human coding of a sample of conversations, so its accuracy is unknown and topic counts may undercount conversations that used vocabulary the scheme did not capture. All topic coding, the keyword matching applied to the AI log and the manual mapping of exam questions to topics, was carried out solely by the author, with no second coder independently verifying these assignments. The comparison between sources was kept descriptive rather than inferential, given only six topic categories, so the patterns reported here should be read as a starting point for further testing, not as confirmed. Three exam questions with genuinely relevant content could not be used in the comparison because their scores mixed several constructs into one number. This remains a single course at a single institution, with 58 students in the AI-log sample after cleaning, so how far these specific topic-level findings generalize elsewhere is untested.

Future Research

The practical takeaway for anyone teaching a similar course is that an AI interaction log and a classroom response system appear to capture different things, and relying on only one may leave a topic like point estimation undetected that the other would have caught. A practical starting point is to treat AI-log volume as a signal of where student attention concentrates, not of where students are struggling, and to read it alongside classroom-response error rates rather than in place of them; where the two disagree, as they did here for point estimation, that disagreement is itself the signal worth following up on. As a working rule for instructors, the two signals can be read together in four cases. High AI-log volume with a high classroom error rate marks a topic worth revisiting. High volume with a low error rate may mark a subtle topic that prompts follow-up questions. Low volume with a high error rate, as for Type I and II errors and statistical power here, may mark a difficulty students are not voicing, and calls for a targeted check. Low volume with a low error rate calls for no action. Where one source has no data on a topic, as for point estimation here, the gap itself is a prompt to add a classroom item. This rule is a hypothesis for future testing, not a finding of this study. For researchers working with AI interaction logs specifically, the preprocessing steps documented here, verifying course scope, checking for non-student accounts, reconstructing broken exports, deserve more explicit attention than they currently receive. A larger sample of AI-tool users would let the topic-level comparison attempted here be tested formally, rather than only described. A version of this study restricted strictly to the three tutorial windows, rather than the full course period, would isolate in-session AI use from the much larger volume of self-directed use captured in this dataset, and would be a natural next step for statistics education research building on this design.

Conclusion

This study compared AI interaction logs, classroom polling data, and final exam records within the same second-year undergraduate statistics course, and found that the three sources point to different, only partly overlapping, pictures of where students struggle. AI-log volume did not consistently align with Wooclap error rates for the topics present in both sources, and point estimation, the topic with the highest exam error rate among directly comparable items, had no Wooclap item at all across three tutorial sessions. These findings are descriptive rather than confirmatory, drawn from a single course, a voluntary AI-tool sample of 58 students, and a six-category topic scheme, so they should be read as a starting point for further work rather than an established result. Within those boundaries, the practical implication is that no single data source captured the full range of student difficulty in this course, and that reading AI interaction logs alongside existing classroom analytics may help to flag topics that either source alone would miss.

Acknowledgment

This research was partially supported by the "Innovate and Share it EEMCS 2025" grant, University of Twente.

Conflict of Interest

The author declares no conflict of interest. Neither the author, the instructional team of the course, nor the University of Twente had any financial, advisory, governance, or developmental relationship with Alexandria or its developers at the time of the study or during the data collection and analysis reported here. A licence agreement between the University of Twente and the platform's developers was established after the study concluded.

References

[1] A. E. Castro Sotos, S. Vanhoof, W. Van den Noortgate, and P. Onghena, "Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education," Educational Research Review, vol. 2, no. 2, pp. 98–113, 2007. https://doi.org/10.1016/j.edurev.2007.04.001

[2] R. Hoekstra, R. D. Morey, J. N. Rouder, and E.-J. Wagenmakers, "Robust misinterpretation of confidence intervals," Psychonomic Bulletin & Review, vol. 21, no. 5, pp. 1157–1164, 2014. https://doi.org/10.3758/s13423-013-0572-3

[3] F. Lyman, "The responsive classroom discussion: The inclusion of all students," Mainstreaming Digest, University of Maryland, College Park, MD, pp. 109–113, 1981.

[4] J. McTighe and F. T. Lyman, "Cueing thinking in the classroom: The promise of theory-embedded tools," Educational Leadership, vol. 45, no. 7, pp. 18–24, 1988.

[5] A. R. Corbo and D. G. G. Sasaki, "Using active learning methodologies in an introductory statistics course," Statistics Education Research Journal, vol. 20, no. 2, Art. 2, 2021. https://doi.org/10.52041/serj.v20i2.113

[6] K. M. Cooper, J. N. Schinske, and K. D. Tanner, "Reconsidering the share of a think-pair-share: Emerging limitations, alternatives, and opportunities for research," CBE—Life Sciences Education, vol. 20, no. 1, fe1, 2021. https://doi.org/10.1187/cbe.20-08-0200

[7] M. K. Smith et al., "Combining peer discussion with instructor explanation increases student learning from in-class concept questions," CBE—Life Sciences Education, 2011. https://doi.org/10.1187/cbe.10-08-0101

[8] B. Catala, D. Muñoz-Higueras, and J. Sanjuán, "Student Response Systems in higher education: A comparative analysis using Wooclap platform in economic courses," Journal of Management and Business Education, 2024. https://journaljmbe.com/article/view/6398

[9] I. Moreno-Medina, M. Peñas-Garzón, C. Belver, and J. Bedia, "Wooclap for improving student achievement and motivation in the Chemical Engineering Degree," Education for Chemical Engineers, vol. 45, pp. 11–18, 2023. https://doi.org/10.1016/j.ece.2023.07.003

[10] A. Dange and R. E. Lopez, "Chatbot Conversations in Physics Education: Using Artificial Intelligence to Analyze Student Reasoning through Computational Grounded Theory," arXiv:2603.04616, 2026. https://arxiv.org/pdf/2603.04616

[11] S. Gok and R. L. Goldstone, "How do students reason about statistical sampling with computer simulations? An integrative review from a grounded cognition perspective," Cognitive Research: Principles and Implications, vol. 9, Art. 33, 2024. https://doi.org/10.1186/s41235-024-00561-x

[12] A. Reinhart, C. Evans, A. Luby, J. Orellana, M. Meyer, J. Wieczorek, P. Elliott, P. Burckhardt, and R. Nugent, "Think-aloud interviews: A tool for exploring student statistical reasoning," Journal of Statistics and Data Science Education, vol. 30, no. 2, pp. 100–113, 2022. https://doi.org/10.1080/26939169.2022.2063209

[13] A. Al-Abri, "Exploring ChatGPT as a virtual tutor: A multi-dimensional analysis of large language models in academic support," Education and Information Technologies, vol. 30, pp. 17447–17482, 2025. https://doi.org/10.1007/s10639-025-13484-x

[14] L. Al Labadi and A. Ly, "Enhancing statistics education through project-based learning (PBL) and the emergence of ChatGPT," Teaching Statistics, vol. 47, no. 3, pp. 200–218, 2025. https://doi.org/10.1111/test.12405

[15] A. R. Ellis and E. Slade, "A new era of learning: Considerations for ChatGPT as a tool to enhance statistics and data science education," Journal of Statistics and Data Science Education, vol. 31, no. 2, pp. 128–133, 2023. https://doi.org/10.1080/26939169.2023.2223609

[16] S. Colando, E. Franke, G. Weinberg, and A. Reinhart, "Analyzing Students' Statistics Writing Before and After the Emergence of Large Language Models," arXiv:2606.22735, 2026. https://arxiv.org/pdf/2606.22735

[17] L. S. DeLuca, A. Reinhart, G. Weinberg, M. Laudenbach, S. Miller, and D. W. Brown, "Developing Students' Statistical Expertise Through Writing in the Age of AI," Journal of Statistics and Data Science Education, vol. 33, no. 3, pp. 266–278, 2025. https://doi.org/10.1080/26939169.2025.2497547

[18] Odea (formerly Library of Alexandria). [Online]. Available: https://odea.education/ (accessed Oct. 9, 2026).