

SATRIA, an adaptive statistics application with limited AI assistance: Development and effects on vocational students' numeracy and motivation

Normansyah, A.B.^{1*}, Yaniawati, P.²

¹Normansyah, A. B./Universitas Pasundan, Jawa Barat, Indonesia

²Yaniawati, P/ Universitas Pasundan, Jawa Barat, Indonesia

*Email: alif.bn@gmail.com

Abstract

Vocational students often meet statistics as a set of formulas, yet their own field asks them to read network data, compare services, and judge claims. This study developed SATRIA, a web-based statistics application for Grade XI students of Computer and Telecommunication Network Engineering, and examined changes in numeracy and learning motivation. SATRIA combines mastery learning, rule-based adaptive pathways, and AI hints limited to five per task that never reveal final answers. An embedded mixed methods design was used: the application was developed in three stages, reviewed by three content and three media experts, and tried out in a nonequivalent control group quasi-experiment with 32 experimental and 30 control students over six meetings. Experts supported its validity (content 94.95%, media 84.76%), and students rated it very practical (84.32% in the limited trial, 89.20% in the main use). Mean numeracy change was 54.30 points in the SATRIA class and 35.83 in the regular class, $t(60) = 5.905$, $p < .001$, $d = 1.50$. Motivation rose in both classes without a reliable difference between them ($p = .389$). Interviews and student work pointed to clearer learning pathways and better data reasoning, alongside calls for smoother network access, simpler hints, and teacher guidance.

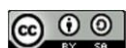
Keywords: mastery learning, adaptive learning, numeracy, learning motivation, vocational education

How to Cite: ¹Normansyah, A. B., ²Yaniawati, P. (2026). SATRIA, an adaptive statistics application with limited AI assistance: Development and effects on vocational students' numeracy and motivation. *Pasundan Journal of Mathematics Education*, x(x), xx-xx. <http://doi.org/10.23969/pjme.v16ix.xxxx>

INTRODUCTION

Numeracy is easy to underestimate. It is tempting to treat it as arithmetic done a little faster, but the PISA framework describes something wider: the capacity to reason mathematically and to formulate, employ, and interpret mathematics in real situations (OECD, 2023). A student who can compute a mean but cannot say what it hides about the stability of a network service has only half of the competence. For students in vocational high schools the stakes are practical as well as academic, because they will later read logs, compare equipment, and choose between options on the strength of data.

Indonesia's recent results show how far this competence still has to travel. In the Programme for International Student Assessment (PISA) 2025, Indonesian 15-year-olds averaged 364 points in



© the Author(s) 2026

This work is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-sa/4.0/)

[doi http://doi.org/10.23969/pjme.v16ix.xxxx](http://doi.org/10.23969/pjme.v16ix.xxxx)



mathematics against an OECD average of 463, and only 17% reached at least Level 2, compared with 65% across the OECD (OECD, 2026). The picture for vocational schools is more encouraging but unfinished. The share of SMK/MAK students reaching the minimum numeracy competency rose from 50.6% in 2022 to 55.1% in 2023 and 57.61% in 2024 (Pratiwi et al., 2025), which still leaves roughly two in five below the minimum. The school where this study took place followed a similar path. Its Education Report moved the numeracy category from low in 2023 to medium in 2024 and good in 2025, while the quality of learning stayed at the medium level in all three reports and scored lower in 2025 than in 2024 (Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi [Kemendikbudristek], 2023, 2024; Kementerian Pendidikan Dasar dan Menengah [Kemendikdasmen], 2025).

Statistics is a natural place to work on this gap, and Computer and Telecommunication Network Engineering (Teknik Jaringan Komputer dan Telekomunikasi, TJKT) is a natural place to do it. Response times, packet loss, signal strength, and transfer durations are everyday numbers for these students, and the Phase E mathematics outcomes ask for exactly the skills needed to make sense of them: determining quartiles and the interquartile range, drawing and interpreting box plots, choosing data representations, examining the relationship between two numerical variables with scatter plots, and evaluating statistical reports (Badan Standar, Kurikulum, dan Asesmen Pendidikan [BSKAP], 2025). Work with real problems has also been reported to support Grade XI students' understanding of basic statistics (Naikofi et al., 2026). There is a curricular wrinkle, however. Phase E belongs to Grade X, whereas Grade XI sits in Phase F (BSKAP, 2025), and the students in this study had not been taught the Phase E statistics outcomes when they were in Grade X. Teaching them in Grade XI was therefore a matter of recovery, and the pretest showed how much recovery was needed: on a 0–100 scale the two participating classes averaged only 26.64 and 24.58.

A gap of that size rarely closes evenly. Some students will have met parts of the content elsewhere, while others need the basics rebuilt. When everyone receives the same sequence, difficulty, and kind of help, the first group waits and the second group moves on without having understood. Mastery learning offers a way out by tying progress to demonstrated competence. Students study a unit, take a formative assessment, receive corrective work or enrichment depending on the result, and are reassessed before moving on (Bloom, 1968). How mastery-based designs relate to progress and learning time, however, has to be examined in each setting (Vlădescu, 2023). Running this cycle by hand for a whole class is demanding, which is where adaptive systems become attractive. Research on adaptive learning indicates that outcomes depend on the accuracy of diagnosis, the quality of feedback, and the fit between the follow-up activity and the learner's need (Wang et al., 2023; Gligorea et al., 2023), and a meta-analysis of artificial intelligence (AI) in education found benefits for achievement and learning perception when AI sits inside a clear pedagogical design (Zheng et al., 2023).

Adding generative AI to this picture invites enthusiasm and worry in equal measure. Large language models can personalize feedback, but they also bring risks of inaccuracy, dependence, and reduced cognitive engagement (Kasneci et al., 2023), and a study in high school mathematics found that AI without guardrails can harm learning, particularly when students can obtain answers directly (Bastani et al., 2025). The design question is therefore less whether to use AI than where to put it and how tightly to bound it.

SATRIA (Statistika Adaptif, Terarah, Responsif, dan Interaktif berbasis AI) was developed with these considerations in mind. It is a web application for Grade XI TJKT statistics that joins three ideas. Mastery learning decides what students must achieve before moving on, rule-based adaptive pathways decide what to do next, and AI offers stepwise hints when a student is stuck. Scoring, mastery decisions,

and the choice of learning pathway stay with rules set by the researcher and open to the teacher's inspection, so the AI never grades and never hands over the answer.

Learning gains are only one side of the story. Students who find the material relevant, understand what they are working toward, and experience success tend to stay engaged longer, whereas repeated failure wears down attention and confidence. The ARCS model (attention, relevance, confidence, satisfaction) offers a vocabulary for designing and examining such experiences (Fang et al., 2024), and motivation has been linked with mathematics achievement and numeracy in earlier studies (Iyamuremye et al., 2023; Lestari et al., 2024). In this study numeracy and motivation were therefore treated as separate outcomes, so that a gain in one would not be mistaken for a gain in the other.

Earlier work has supplied many of the building blocks. Statistics e-learning materials have been developed for vocational students (Supianti et al., 2022), mobile digital materials have been paired with mathematical literacy (Yaniawati et al., 2023), adaptive digital books have been built for numeracy (Yerizon et al., 2025), adaptive digital assessment has been explored in this journal (Santika et al., 2026), and generative AI has been combined with ethnomathematics in geometry (Mariyati et al., 2026). Little published work seems to couple a vocational context, explicit mastery rules, and a deliberately capped AI helper in a single statistics package, and then judge that package by both measured outcomes and students' own accounts of using it. This study aimed to address that gap. It described the development of SATRIA, assessed its validity and practicality, compared changes in numeracy and in learning motivation between a SATRIA class and a regular class, and interpreted the effectiveness of its implementation by combining quantitative and qualitative evidence. Two hypotheses guided the comparison: the mean change in numeracy (H1) and in motivation (H2) would differ between students who used SATRIA and students who received regular instruction.

METHODS

Research Design

The study followed an embedded mixed methods design in which quantitative data were primary and qualitative data played a supporting role, usually written QUAN(qual) (Creswell & Plano Clark, 2018). Three strands were nested in one plan. First, research and development produced and refined the application, following the Alessi and Trollip (2001) model of planning, design, and development, with standards, evaluation, and project management running alongside; media built on technology develops through repeated checking and revision of this kind (Alessi & Trollip, 2001; Yaniawati & Indrawan, 2024). Second, a quasi-experiment with a nonequivalent control group and pretest-posttest measurement compared two intact classes. Third, qualitative data (expert comments, revision notes, open-ended answers, and interviews) explained development decisions and the experiences behind the scores. Because students were not randomly assigned, the design limits causal interpretation, and conclusions are restricted to the classes, content, and period studied.

Participants and Setting

The study was conducted at SMKN 1 Cijulang, a public vocational high school in West Java, Indonesia, in the odd semester of the 2026/2027 academic year. The pretest was administered on 6 August 2026 and the posttest on 17 September 2026. The practicality questionnaire for the experimental class followed on 24 September, and interviews took place between 28 and 30 September 2026. The accessible population was all Grade XI TJKT students at the school. Two intact classes were chosen

purposively, taking into account the shared study program, comparable history of statistics instruction, schedule availability, and readiness to take part. Class XI TJKT 2 ($n = 32$) used SATRIA and Class XI TJKT 1 ($n = 30$) received regular instruction. Only students with both pretest and posttest data were analyzed. Sample size was determined by the availability of two whole classes rather than by a power analysis (Lakens, 2022), so the precision of the estimates and the reach of any generalization are limited.

Three content experts and three media experts evaluated the product. A limited trial involved ten Grade XII TJKT 2 students outside the main sample. Eight students from the experimental class (coded E02, E04, E10, E13, E14, E15, E24, and E32) were purposively selected for interviews to reflect variation in numeracy change, motivation change, and use of AI hints. Student data were coded, the key linking codes to identities was stored separately, and interview excerpts are reported with codes only. Participation was voluntary and followed the research permissions of the university and the school; it was not tied to report-card grades or to access to learning services. No personal data or credentials were entered into the development instructions or the AI help.

The SATRIA Application and the Learning Conditions

SATRIA is a web application that arranges Phase E statistics content into learning objectives with explanations, examples, and exercises set in TJKT contexts such as access point latency, packet loss, and internet speed. It also provides formative assessment, feedback, corrective learning, enrichment, and progress records. Its adaptive behavior follows a cycle of diagnosis, decision, intervention, and re-evaluation, governed by rules that the researcher and the teacher can inspect. A unit counts as mastered at a score of 80% or above. Students below the threshold receive remedial material, corrective practice, or reassessment, whereas those at or above it move on or take enrichment tasks. Scoring relies on answer keys, rubrics, and rules defined by the researcher.

When a student struggles, the AI offers help in up to five graded levels per task. A hint can point to relevant information, recall a concept, or suggest the next step, but it never discloses the final answer, assigns a score, or decides mastery. If the hints are not enough, the student is directed to the teacher. OpenAI Codex supported the programming under the researcher's instructions. Content, items, rubrics, adaptive rules, the mastery threshold, and the hint levels were all specified by the researcher, and every technical output was inspected and tested before it became part of the product.

Both classes covered the same learning objectives and content in six meetings of equal length, and the teacher accompanied both. Regular instruction consisted of teacher explanation, examples, practice, discussion, and feedback. The SATRIA class received the same teacher-led components and, in addition, used SATRIA for content, practice, formative assessment, and follow-up activities. Because SATRIA was implemented as a whole package, any difference between the classes is attributed to the package and not to the AI component alone.

Instruments

Numeracy was measured with a package of five essay items set in TJKT contexts: configuration transfer time, network latency, packet loss percentages, the relationship between the number of connected devices and latency, and an internet speed advertisement. The items called for formulating, employing, and interpreting and evaluating mathematics (OECD, 2023; Purnomo et al., 2022). Students had 90 minutes. Each item carried a maximum of four points, so the package maximum was 20 points, and raw scores were converted to a 0–100 scale by dividing by the maximum and multiplying by 100. An

analytic rubric assessed steps and reasoning, so that computation slips and conceptual errors were not treated alike. The same package and rubric were used at pretest and posttest in both classes.

Content suitability was judged by experts and summarized with Aiken's V (Aiken, 1985), shown in Equation (1).

$$V = \frac{\sum s}{n(c-1)} \quad (1)$$

where $s = r - l_0$, r is an expert's rating, l_0 is the lowest rating category, c is the number of rating categories, and n is the number of experts. Expert feedback led to completing the illustrations of several items before use. An empirical try-out with 35 students outside the main sample gave corrected item-total correlations of .494 to .639, difficulty indices of .454 to .618, discrimination indices of .403 to .583, and Cronbach's alpha of .794. All five items were retained. Inter-rater agreement for the rubric was not computed, a limitation taken up later.

Learning motivation was measured with a 24-item questionnaire based on the four ARCS elements (Fang et al., 2024). Each element had six items, three positive and three negative, answered on a five-point agreement scale. Negative items were reverse-scored, so total scores ran from 24 to 120. The items referred to learning statistics in general and not to SATRIA, which allowed both classes to answer the same construct, and statements about navigation or the interface were left to the practicality questionnaire. In the try-out ($n = 35$) corrected item-total correlations ranged from .364 to .641 and alpha was .907. In the main data alpha ranged from .897 to .949 across classes and measurement times. A separate empirical validation of the ARCS structure was not carried out.

Practicality was assessed with a 25-item questionnaire of positive statements on a five-point scale, covering clarity of instructions, navigation and legibility, technical smoothness, adaptive flow and mastery, feedback and AI help, and efficiency and acceptance. Four open questions asked about helpful features, obstacles, experience with AI hints, and suggestions. Experts used five-point rating sheets with space for comments: content experts examined concepts, items, rubrics, language, and the fit between activities and assessment, whereas media experts examined presentation, navigation, interaction, and function (Yerizon et al., 2025). A semi-structured interview guide covered the materials, corrective learning, reassessment, hints, the relevance of the TJKT context, solution strategies, attention, confidence, satisfaction, and teacher support. An alpha of .70 served as the working threshold for internal consistency (Tavakol & O'Brien, 2022).

Data Analysis

Expert and student ratings were converted to percentages by dividing the obtained score by the maximum score and multiplying by 100. Practicality was classified in five 20-point bands: more than 80% to 100% as very practical, more than 60% to 80% as practical, more than 40% to 60% as fairly practical, more than 20% to 40% as less practical, and 0% to 20% as not practical. The five-category structure was adapted from Rohmah et al. (2025), and the interval limits were set by the researcher. For each student the change score was calculated as in Equation (2), separately for numeracy and motivation.

$$\Delta_i = X_{\text{post},i} - X_{\text{pre},i} \quad (2)$$

Mean change scores of the two classes were compared with an independent samples t -test (two-tailed, $\alpha = .05$) after checking normality with the Shapiro–Wilk test and equality of variances with Levene's test. The equal-variances row was read when Levene's p was at least .05 (Navarro & Foxcroft, 2025). Mean differences, standard errors, 95% confidence intervals, and Cohen's d (the mean difference divided

by the pooled standard deviation) were reported, and baseline equivalence was checked with *t*-tests on pretest scores. Item-level means were examined descriptively. A posttest score of at least 80 was set as an operational criterion, distinct from unit mastery inside the application. Pretest scores were not entered as a covariate.

Qualitative data were analyzed through reflexive thematic analysis (Byrne, 2022). Transcripts were read, coded, and interpreted by the researcher, with NVivo used to organize codes, compare themes across interviewee attributes, and produce visualizations. Trustworthiness relied on comparing sources (interviews against open answers and score profiles, expert comments against revision notes), keeping a traceable coding trail, and paying attention to cases that departed from the general pattern. Quantitative and qualitative findings were then placed side by side in joint displays to interpret the effectiveness of implementation (Creswell & Plano Clark, 2025).

RESULTS AND DISCUSSION

Development of SATRIA

Planning began with the need to recover Phase E statistics for Grade XI students and to give the content a vocational face. The material was organized around handling data, choosing diagrams, reading spread, interpreting relationships, and checking claims against data, using network situations students already know. The roles of the three actors were also settled at this point. The application manages activities according to assessment results and the mastery rule, the AI gives hints on request, and the teacher steps in when hints are not understood or when technical problems block the work. At the design stage the content was arranged into explanations, examples, exercises, formative assessments, corrective activities, and enrichment. Each unit carried the 80% mastery threshold, the AI help was designed with up to five hint levels, and the interface showed learning objectives and status so that students could see what they had finished and what was left.

During development the design became a working web application, with Codex supporting the programming, and feedback from experts and users led to several revisions. The opening page, which at first introduced a student workspace and teacher accompaniment, was redesigned to show the SATRIA identity together with network illustrations and data displays, so that the subject matter was visible as soon as the application opened. On the learning page, subtopics had been listed in one long stretch. They were reorganized into panels that open and close, grouped by learning objective and showing mastery status (Figure 1). Illustrations of several items were clarified, and the practice exercises were reviewed to give them more variety.

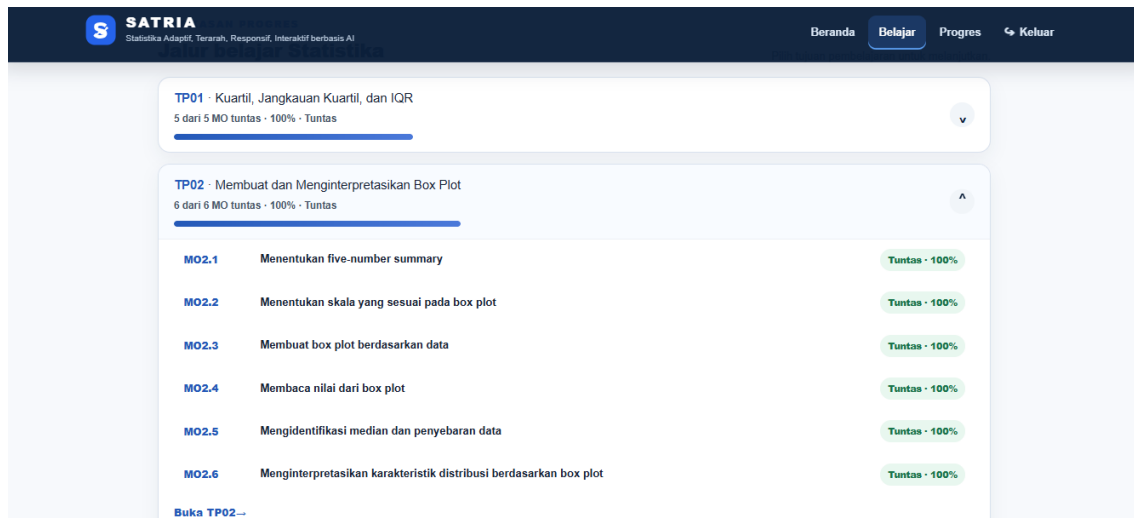


Figure 1. SATRIA Learning Page After Revision, Showing Objective Panels and Mastery Status (Interface in Indonesian)

Classroom use exposed a second kind of problem. One student reported that help appeared only at the beginning of a task (E32), another saw an “assessment unavailable” message (E14), and a third noticed that the next activity opened before the previous one was completed (E10). These reports prompted checks on how assessment, hints, answer saving, and the unlocking of activities interact. They also show why expert approval alone cannot settle whether an application works under real classroom conditions.

In this respect the Alessi and Trollip (2001) model did what it promises. It kept learning needs at the center of development and turned them into content, assessment, corrective activities, and hints that depend on one another. The choice of network contexts gave students a reason to care about statistics, which is consistent with the argument that linking mathematics to students’ vocations helps them see its use (Muhрман, 2022). A familiar context did not guarantee conceptual understanding, though. Two students still had trouble locating quartiles and telling measures of spread apart despite working with network data. Context seems to clarify why a concept is used, while understanding the concept itself still needs examples, practice, and talk about reasoning. The adaptive pathway followed the logic of using learning data to adjust activities (Gligorea et al., 2023), but the experience of E04, who at first took a hand-off to the teacher for a system fault, shows that on-screen messages must explain what happens next. Finally, using Codex for programming while leaving instructional decisions with the researcher matches the call for human control over AI in education (Miao & Holmes, 2023). A product that can be generated quickly still has to be checked against the content, the rules, and the people who use it.

Validity and Practicality of SATRIA

Table 1 summarizes the judgments of experts and students. The content experts awarded 470 of 495 possible points (94.95%, mean rating 4.75, Aiken’s $V = .937$), and the media experts awarded 356 of 420 (84.76%, mean rating 4.24, $V = .810$). Both coefficients lie close to 1, with stronger agreement on content than on media. Comments steered the revisions: remarks on illustrations and exercise variety concerned clarity and opportunities to practice, while remarks on layout and navigation concerned ease of use. The decision to use the application therefore rested on the scores together with the comments and the follow-up revisions, not on the percentages alone.

Table 1. Summary of Expert Validation and Student Practicality Results

Source	<i>n</i>	Score	Maximum	Percentage	Aiken's <i>V</i>
Content experts	3	470	495	94.95%	.937
Media experts	3	356	420	84.76%	.810
Students, limited trial	10	1,054	1,250	84.32%	–
Students, main use	32	3,568	4,000	89.20%	–

Students judged SATRIA very practical in both trials (Table 2). The limited trial yielded 1,054 of 1,250 points (84.32%), and the main use 3,568 of 4,000 (89.20%). Because the two trials involved different groups and a revised product, the higher figure in the main use is not read as a change of opinion among the same students. Feedback and AI help earned the highest percentage in the main use (91.50%), which suggests that students found the hints and feedback supportive of their work. Efficiency and acceptance scored lowest (87.34%), and student comments tied this to loading time, connection, devices, and answer submission. Open-ended responses named the concise material, graphs, assessments, and mastery information as helpful, and the requests included automatic saving, a learning history, more flexible access, video, and more varied activities.

Table 2. Practicality Percentages by Aspect

Aspect	Limited trial (<i>n</i> = 10)	Main use (<i>n</i> = 32)
Clarity of instructions	85.50%	88.28%
Navigation and legibility	82.00%	88.75%
Technical smoothness	83.00%	88.44%
Adaptive flow and mastery	87.50%	90.31%
Feedback and AI help	85.20%	91.50%
Efficiency and acceptance	82.50%	87.34%
Overall	84.32%	89.20%

The numbers do not tell the whole story. E15 described the application as smooth in the questionnaire but mentioned a slow device and unsaved answers in the interview. E24 gave a short answer about being able to ask for help yet explained that AI help had not been used. Reading the two sources together gave a more accurate picture than either alone. Cronbach's alpha for the practicality questionnaire was .697 in the limited trial, slightly below the .70 working threshold, and .968 in the main use. The first value is reported as it is rather than rounded up, and with only ten participants it is unstable in any case.

Practicality in this study meant more than usability on paper. Rohmah et al. (2025) stress digital learning that matches the needs of vocational students, and the suggestions about connection, saving, and variety are exactly that kind of concrete need. When answers were lost, E14 explained, work had to be repeated, time ran short, and activities became rushed, and slow-loading hints weakened the help in the same way. Language mattered too. E15 hesitated to respond to hints for fear of being wrong, and E04 needed a more directed explanation, a reminder that a hint on the screen helps only when the student can understand it. Teachers remained important for orienting students to the flow, receiving their difficulties, and handling technical obstacles, and E02 grasped how to continue only after using the application several times. SATRIA can therefore be called practical when it comes with orientation, adequate connectivity, and reliable saving.

Group Comparisons of Numeracy and Motivation Change

Table 3 presents descriptive statistics for both outcomes. The classes began at virtually the same place. Pretest numeracy means were 26.64 (SATRIA) and 24.58 (regular), $t(60) = 0.606$, $p = .547$, and pretest motivation means were 73.56 and 73.50, $t(60) = 0.020$, $p = .984$. Failing to detect a difference does not make the classes identical, since they were intact groups. At posttest the SATRIA class averaged 80.94 in numeracy and the regular class 60.42, while motivation averaged 99.53 and 96.73 (Figure 2). Every student in both classes showed a positive change on both outcomes.

Table 3. Descriptive Statistics of Numeracy and Motivation Scores

Outcome	Class	<i>n</i>	Pretest <i>M</i> (<i>SD</i>)	Posttest <i>M</i> (<i>SD</i>)	Change <i>M</i> (<i>SD</i>)
Numeracy (0–100)	SATRIA	32	26.64 (13.80)	80.94 (14.04)	54.30 (14.24)
	Regular	30	24.58 (12.85)	60.42 (12.58)	35.83 (9.81)
Motivation (24–120)	SATRIA	32	73.56 (11.79)	99.53 (10.34)	25.97 (12.19)
	Regular	30	73.50 (13.09)	96.73 (13.95)	23.23 (12.61)

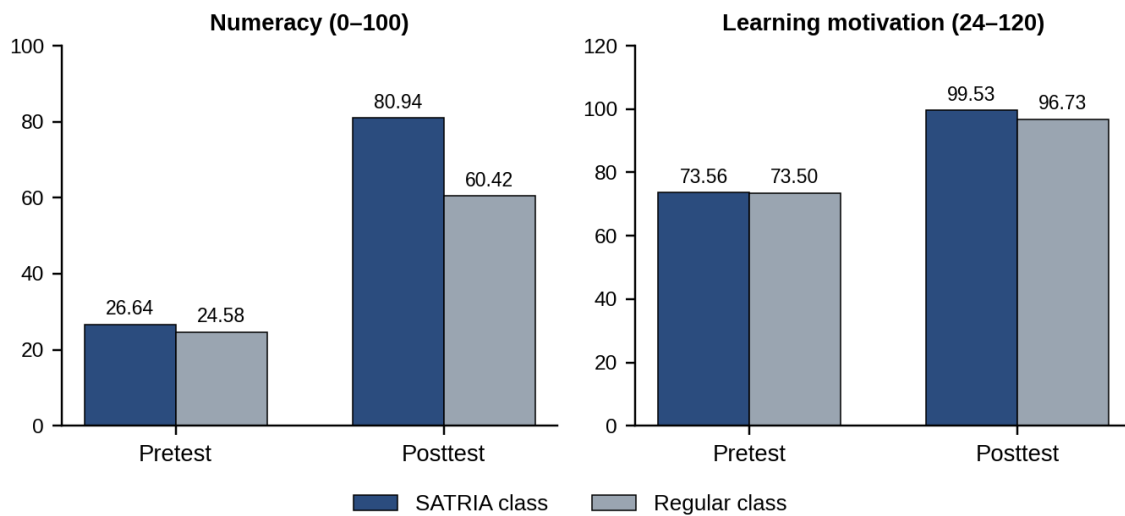


Figure 2. Pretest and Posttest Means of Numeracy and Motivation by Class

Change scores met the assumptions for a *t*-test. For numeracy, Shapiro–Wilk tests gave $p = .171$ (SATRIA) and $p = .412$ (regular), and Levene's test gave $F = 2.782$, $p = .101$. For motivation the corresponding values were $p = .238$ and $p = .788$, and $F = 0.129$, $p = .721$. Equal variances were therefore assumed (Table 4). Numeracy change was larger in the SATRIA class ($M = 54.30$, $SD = 14.24$) than in the regular class ($M = 35.83$, $SD = 9.81$), a difference of 18.46 points, $t(60) = 5.905$, $p < .001$, 95% CI [12.21, 24.72], $d = 1.50$, so H1 was supported. Motivation change was 25.97 ($SD = 12.19$) against 23.23 ($SD = 12.61$). The difference of 2.74 points was not statistically significant, $t(60) = 0.868$, $p = .389$, 95% CI [−3.56, 9.04], $d = 0.22$, and H2 was not supported. A non-significant result does not show that the two conditions are equal, only that these data do not demonstrate a difference.

Table 4. Independent Samples *t*-Test of Change Scores

Outcome	Mean difference	SE	<i>t</i> (60)	<i>p</i>	95% CI	Cohen's <i>d</i>
Numeracy	18.46	3.13	5.905	< .001	[12.21, 24.72]	1.50
Motivation	2.74	3.15	0.868	.389	[-3.56, 9.04]	0.22

Understanding the Numeracy Gains

The posttest criterion tells a more sobering story than the averages. Seventeen of 32 students in the SATRIA class (53.13%) scored at least 80, compared with one of 30 (3.33%) in the regular class, yet nearly half of the SATRIA class (46.88%) still fell short. At the item level, SATRIA-class posttest means were above 3 out of a possible 4 on all five items. The scatter plot item was highest (3.50) and the quartile and spread item lowest (3.03), whereas regular-class means ranged from 2.15 to 2.70. The gains were therefore real but uneven.

The pattern fits the logic of mastery learning, in which assessment is used to decide on corrective instruction before students move on (Bloom, 1968), and it fits findings that adaptive systems can support learning when adjustments rest on relevant information about the learner (Wang et al., 2023). Two cautions temper this reading. SATRIA was implemented as one package, so the effects of its mastery rules, TJKT context, hints, and teacher accompaniment cannot be separated, and the AI component in particular cannot be credited on its own. In addition, only one class served in each condition, so class characteristics are entangled with the treatment, and the same test was used twice, so familiarity with the items may have contributed to the large gains in both classes.

Students' written work shows what the gains looked like. On the packet loss task, students had to choose a representation, count the tests with a loss of no more than 3%, and judge whether a 75% target had been met. Ten of sixteen tests qualified, which is 62.5%, so the target was not met. At pretest, E02 chose a box plot simply because the shape was familiar and counted only values below 3%. At posttest the same student picked a dot plot so that every test and its frequency stayed visible, included the values at the 3% boundary, obtained 62.5%, and concluded that the target was missed (numeracy score 17.50 to 90.00). The wording matters here: "no more than" decides which values count, and a student who watches only the arithmetic can miss it.

E13's growth lay in evaluating claims. An advertisement boasted of speed three times faster when the figure had moved from 80 to 84 Mbps. At pretest E13 trusted the height of the bars and divided by the final value. At posttest the student used the starting value, found an increase of 5%, rejected the claim, pointed out that the axis began at 78 and exaggerated the gap, and added that stability and subscription cost deserved consideration (25.00 to 100.00). A student from the regular class (K23) showed that progress was not exclusive to SATRIA: after instruction this student stopped treating a scatter plot of device count and latency as proof of cause and named signal interference and bandwidth capacity as other possible factors. Taken together, these examples correspond to the formulate-employ-interpret cycle of the PISA framework (OECD, 2023). Choosing a representation is formulation, reading a threshold is employing, and weighing a claim against the evidence is interpretation and evaluation.

Other students showed what remained to be done. E04 (17.50 to 55.00) moved from a histogram to a dot plot but left the axis without plotted points and still counted only values below 3%. E10 (0.00 to 67.50) drew a correct dot plot with the right frequencies but wrote 6.25% instead of 62.5%. E15 (22.50 to 47.50) still treated the quartiles as the minimum and maximum and mixed up the range and the

interquartile range. E24 (47.50 to 87.50) computed 62.5% correctly but drew one marker at 8% where the data had two. These slips differ in kind, conceptual in one case, computational in another, and a matter of unfinished work in a third, and each calls for a different corrective response. Repeating the same task without clarifying the source of difficulty, as E04 and E15 experienced it, tires students more than it helps them. A habit of asking whether a result is reasonable given the number of data points would have caught E10's error, and comparing what the range and the interquartile range each summarize in a single data set could help E15. Unfinished answers should be investigated before they are labeled misconceptions.

Understanding the Motivation Results

Motivation rose in both classes, but the rise did not differ reliably between them. Table 5 shows that all four ARCS elements increased in both classes. Confidence showed the largest change (8.19 in the SATRIA class and 7.20 in the regular class), followed by satisfaction (6.94 and 6.23), and attention and relevance also rose. These element scores are descriptive and were not tested separately. The pattern suggests that what changed was not only interest in the activities but also students' sense that they could face the tasks and had succeeded at them.

Table 5. Mean Scores of the ARCS Elements and Their Change (Range 6–30, Descriptive)

ARCS element	SATRIA: pretest → posttest	Change	Regular: pretest → posttest	Change
Attention	18.94 → 24.50	5.56	18.83 → 24.03	5.20
Relevance	19.63 → 24.91	5.28	19.47 → 24.07	4.60
Confidence	16.84 → 25.03	8.19	16.87 → 24.07	7.20
Satisfaction	18.16 → 25.09	6.94	18.33 → 24.57	6.23

Interviews added texture. Attention was caught by graphs and network contexts: E02 was helped by the box plot illustration, and E14 recognized the use of latency in the school's own network. It slipped when texts were long or hints loaded slowly. Relevance helped students see why statistics mattered, but relevance alone did not sustain effort when a concept stayed hard or the activities felt repetitive, in line with the view that motivation depends on experiences of progress (Fang et al., 2024). Confidence grew where students recognized a step they could take. E10 became more willing to try and E02 saw that early difficulties could be repaired, while E15 remained afraid of making mistakes. Satisfaction came from finishing units, and E04 felt relieved on completing activities. E13 and E24 asked for further challenges, since repeated activities can dull the engagement of students who have already mastered the content.

That the SATRIA class did not out-gain the regular class in motivation should not come as a surprise. Regular lessons also provide explanation, practice, interaction with the teacher, and the experience of success, so rising motivation in the regular class may simply reflect that support. Motivation and ability also moved at different speeds. E10 grew bolder while some skills remained weak, and E24 produced strong numeracy work with only a modest motivational change. Because motivation was self-reported and the structure of the questionnaire had not been validated separately, these results are read with caution.

Students' Use of AI Hints and the Role of the Teacher

How students used the hints proved as informative as whether they used them. E02 asked for help to understand why data must be ordered first, and E32 described the hints as giving direction rather

than answers: “it gives hints, not answers” (translated from Indonesian). E10 deliberately entered wrong answers to coax more specific hints, a strategy that shows ingenuity but also a risk, that of getting close to the answer without doing the work of solving. E24 preferred to retry alone and reported not using the AI at all, yet finished with strong results, and E15 was reluctant to respond to hints. The interview matrices placed solution strategies, the sufficiency of the five-hint limit, and hand-offs to the teacher among frequent users, and reasons for avoiding AI among occasional and non-users. The counts in these matrices show how codes and attributes co-occur. They are not numbers of students, and they do not show that more frequent use led to larger gains.

These accounts echo cautions in the literature. Kasneci et al. (2023) note both the opportunities and the challenges of language models in education, and Bastani et al. (2025) separate success while AI is available from mastery once it is withdrawn. E10’s baiting of the hints is a reminder to inspect what hints say and how students respond to them, and to pair the five-hint cap with follow-up that keeps students thinking. E24’s experience indicates that students who can work independently should be free to do so. Support should fit the person and should not oblige everyone to use every feature.

Teacher support remained indispensable. E32 put it plainly: “I need to discuss it, sir, not just get hints” (translated). Switching to the teacher appeared in every interviewee profile, including students who did well on both outcomes. Technical problems such as internet connection and data saving appeared across several profiles, and overly long or complex explanations were mentioned by students with low results on both outcomes. In line with Miao and Holmes (2023), the teacher connects the hint on the screen with the difficulty a student is actually having and can ask the student to explain the reasoning behind a step. The qualitative themes also showed that an adaptive pathway works best when students understand it. E13 appreciated being tested first, in the student’s words, “I get tested first, so what I already know I don’t have to study again” (translated), whereas E04 and E15 found the repetition tiring and E24 wanted more varied tasks after mastering the content.

Integrated Interpretation and Limitations

Bringing the strands together, the evidence supports SATRIA most clearly as a way of organizing learning and strengthening numeracy under the conditions studied. Experts supported its content and media quality, students found it very practical, numeracy change was substantially larger than in regular instruction, and student work showed better selection of representations, reading of thresholds, and evaluation of claims. The evidence for motivation is weaker. Both classes improved, no difference was shown, and students’ experiences ranged from growing courage to fear of mistakes to a need for harder tasks. The qualitative evidence also located what needs improving: more precise diagnosis of difficulty, simpler language in the hints, richer enrichment tasks, reliable saving and loading, and sustained teacher accompaniment. Because the whole package was tested, none of these gains can be assigned to a single feature, least of all to the AI.

Several limitations qualify these conclusions. With one class per condition in one school, class characteristics are inseparable from the treatment, and an analysis of change scores without covariates does not control for unmeasured factors such as study habits outside class. Using the same test twice may have inflated familiarity with the item formats, and no inter-rater coefficient was reported for the numeracy rubric. The motivation questionnaire relied on self-report without separate construct validation, and the limited trial used Grade XII students and produced an alpha slightly under the working threshold. The in-depth analysis of written work covered seven cases, so it cannot estimate how common each kind of error was. Interviews came only from the SATRIA class, and the researcher acted as both teacher and

interviewer, which may have shaped the responses. Data on AI use came mainly from interviews and open answers instead of interaction logs, and no delayed posttest was given, so the durability of gains in numeracy and motivation is unknown.

CONCLUSION

This study developed SATRIA, an adaptive statistics application that joins mastery learning, rule-based pathways, and capped AI hints in a TJKT context, and examined it with an embedded mixed methods design. The application was developed through planning, design, and development with the Alessi and Trollip model and was revised after expert and user input. Experts supported its validity, and students rated it very practical in both the limited trial and the main use. The SATRIA class improved its numeracy more than the regular class (a difference of 18.46 points in mean change, $d = 1.50$), and 17 of 32 SATRIA students reached the posttest criterion of 80 against one of 30 in the regular class. Motivation rose in both classes without a demonstrable difference between them, and students' experiences were mixed. Overall, SATRIA appears effective mainly in structuring learning and strengthening numeracy, as a package and in the classes studied.

Several practical implications follow. Teachers can use assessment results to decide on corrective work or enrichment, but they should also read students' reasoning, since a correct final answer can sit beside a misread boundary and a wrong one beside a sound plan. Developers should give priority to answer saving, clear status messages, faster loading of hints, simpler hint wording, and richer enrichment, and should test on the devices and connections students actually have. Schools should ensure devices, internet access, and orientation time, and keep alternative activities ready for when connectivity fails. AI hints work best as companions that help students interpret information and check their own reasoning, with the teacher close by.

Future research could extend the evaluation to more classes, schools, and longer periods, follow the retention of numeracy and motivation, record hint use through logs linked to students' work, and compare designs with and without AI hints to separate the contribution of each component.

Acknowledgments

The authors thank the students and teachers of SMKN 1 Cijulang for their participation and cooperation, the three content experts and three media experts who reviewed SATRIA, and supervisors and other supporters.

Declarations

- Author Contribution : Author 1: Conceptualization, Writing - Original Draft, Editing and Visualization; Author 2: Writing - Review & Editing, Formal analysis, and Methodology; Author 3: Validation and Supervision (<https://www.elsevier.com/authors/policies-and-guidelines/credit-author-statement>)
- Funding Statement : Funding statement to be completed by the authors (name the funder and grant number, or state that no external funding was received).
- Conflict of Interest : The authors declare no conflict of interest.
- Additional Information : Additional information is available for this paper.

REFERENCES

- Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement*, 45(1), 131–142. <https://doi.org/10.1177/0013164485451012>
- Alessi, S. M., & Trollip, S. R. (2001). *Multimedia for learning: Methods and development* (3rd ed.). Allyn & Bacon.
- Badan Standar, Kurikulum, dan Asesmen Pendidikan. (2025). *Keputusan Kepala BSKAP Nomor 046/H/KR/2025 tentang capaian pembelajaran pada pendidikan anak usia dini, jenjang pendidikan dasar, dan jenjang pendidikan menengah* [Decree of the Head of BSKAP Number 046/H/KR/2025 on learning outcomes for early childhood education and primary and secondary education]. Kementerian Pendidikan Dasar dan Menengah.
- Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakci, O., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26), Article e2422633122. <https://doi.org/10.1073/pnas.2422633122>
- Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment*, 1(2), 1–12.
- Byrne, D. (2022). A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & Quantity*, 56(3), 1391–1412. <https://doi.org/10.1007/s11135-021-01182-y>
- Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.
- Creswell, J. W., & Plano Clark, V. L. (2025). *Designing and conducting mixed methods research* (4th ed.). SAGE Publications.
- Fang, X., Ng, D. T. K., Leung, J. K. L., & Xu, H. (2024). The applications of the ARCS model in instructional design, theoretical framework, and measurement tool: A systematic review of empirical studies. *Interactive Learning Environments*, 32(10), 5919–5946. <https://doi.org/10.1080/10494820.2023.2240867>
- Gligorea, I., Cioca, M., Oancea, R., Gorski, A.-T., Gorski, H., & Tudorache, P. (2023). Adaptive learning using artificial intelligence in e-learning: A literature review. *Education Sciences*, 13(12), Article 1216. <https://doi.org/10.3390/educsci13121216>
- Iyamuremye, E., Ndayambaje, I., & Muwonge, C. (2023). Relationships of mathematics achievement with self-determined motivation and mathematics anxiety among senior two students in Northern Rwanda. *Heliyon*, 9(4), Article e15411. <https://doi.org/10.1016/j.heliyon.2023.e15411>
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, Article 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kementerian Pendidikan Dasar dan Menengah. (2025). *Rapor Pendidikan milik SMKN 1 Cijulang tahun 2025* [Education report of SMKN 1 Cijulang 2025] [Poster].
- Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi. (2023). *Rapor Pendidikan SMKN 1 Cijulang tahun 2023* [Education report of SMKN 1 Cijulang 2023] [Poster].
- Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi. (2024). *Rapor Pendidikan milik SMKN 1 Cijulang tahun 2024* [Education report of SMKN 1 Cijulang 2024] [Poster].
- Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, 8(1), Article 33267. <https://doi.org/10.1525/collabra.33267>

- Lestari, A., Septian, A., & Inayah, S. (2024). The influence of learning motivation and numeracy on high school students. *PRISMA*, 13(1), 81–89. <https://doi.org/10.35194/jp.v13i1.4140>
- Mariyati, I., Darmawan, M. S., & Romeo, N. (2026). Strengthening prospective mathematics teachers' geometry understanding through ethnomathematics integrated with generative artificial intelligence. *Pasundan Journal of Mathematics Education*, 16(1), 12–24. <https://journal.unpas.ac.id/index.php/pjme/article/view/41543>
- Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO.
- Muhrman, K. (2022). How can students in vocational education be motivated to learn mathematics? *Nordic Journal of Vocational Education and Training*, 12(3), 47–70. <https://doi.org/10.3384/njvet.2242-458X.2212347>
- Naikofi, P. A., Siahaan, M. M. L., & Bone, D. (2026). Exploring student's conceptual understanding of statistics towards the problem-based learning model. *Pasundan Journal of Mathematics Education*, 16(1), 74–86. <https://doi.org/10.23969/pjme.v16i1.40739>
- Navarro, D. J., & Foxcroft, D. R. (2025). *Learning statistics with jamovi: A tutorial for beginners in statistical analysis*. Open Book Publishers. <https://doi.org/10.11647/OBP.0333>
- OECD. (2023). *PISA 2022 assessment and analytical framework*. OECD Publishing. <https://doi.org/10.1787/dfe0bf9c-en>
- OECD. (2026). *PISA 2025 results (Volume I): Future-ready students*. OECD Publishing. <https://doi.org/10.1787/73451bc5-en>
- Pratiwi, I., Samosir, I., Ekowati, D. R., Wicaksono, M. E., Krismahardhika, B. G., Rihardika, A., Krisna, F. N., & Martak, Y. F. (2025). *Rapor Pendidikan Indonesia 2025: Capaian dan agenda perbaikan* [Indonesia education report 2025: Achievements and improvement agenda]. Pusat Standar dan Kebijakan Pendidikan.
- Purnomo, H., Sa'dijah, C., Hidayanto, E., Permadi, H., & Anwar, L. (2022). Development of instrument numeracy skills test of minimum competency assessment (MCA) in Indonesia. *International Journal of Instruction*, 15(3), 635–648. <https://doi.org/10.29333/iji.2022.15335a>
- Rohmah, L., Yaniawati, P., Saputra, I., & Sharif, S. (2025). Mobile learning innovation to improve vocational high school students' mathematical communication skills and dispositions. *Pasundan Journal of Mathematics Education*, 15(1), 1–15. <https://doi.org/10.23969/pjme.v15i1.24123>
- Santika, S., Nugraha, D. A., & Mansyur, M. Z. (2026). Adaptive digital assessment of learning styles: Enhancing personalized mathematics instruction through sensory preferences. *Pasundan Journal of Mathematics Education*, 16(1), 52–62. <https://journal.unpas.ac.id/index.php/pjme/article/view/41250>
- Sari, N. M., Yaniawati, P., Supianti, I. I., & Indriani, R. (2025). Digital game-based learning interventions on students' numeracy skills and engagement. *Formatif: Jurnal Ilmiah Pendidikan MIPA*, 15(1), 39–50. <https://doi.org/10.30998/formatif.v15i1.23356>
- Supianti, I. I., Yaniawati, P., Osman, S. Z. M., Al-Tamar, J., & Lestari, N. (2022). Development of teaching materials for e-learning-based statistics materials oriented towards the mathematical literacy ability of vocational high school students. *Infinity Journal*, 11(2), 237–254. <https://doi.org/10.22460/infinity.v11i2.p237-254>
- Tavakol, M., & O'Brien, D. (2022). Psychometrics for physicians: Everything a clinician needs to know about assessments in medical education. *International Journal of Medical Education*, 13, 100–106. <https://doi.org/10.5116/ijme.625f.bfb1>

- Vlădescu, C. (2023). The relationship between mastery learning models and academic achievement in mathematics. *International Electronic Journal of Mathematics Education*, 18(4), Article em0751. <https://doi.org/10.29333/iejme/13705>
- Wang, S., Christensen, C., Cui, W., Tong, R., Yarnall, L., Shear, L., & Feng, M. (2023). When adaptive learning is effective learning: Comparison of an adaptive learning system to teacher-led instruction. *Interactive Learning Environments*, 31(2), 793–803. <https://doi.org/10.1080/10494820.2020.1808794>
- Yaniawati, P., Fisher, D., Permadi, Y. D., & Yatim, S. A. M. (2023). Development of mobile-based digital learning materials in blended learning oriented to students' mathematical literacy. *International Journal of Information and Education Technology*, 13(9), 1338–1347. <https://doi.org/10.18178/ijiet.2023.13.9.1936>
- Yaniawati, R. P., & Indrawan, R. (2024). *Metodologi penelitian: Konsep, teknik, dan aplikasi* [Research methodology: Concepts, techniques, and applications]. Refika Aditama.
- Yerizon, Arnellis, Suherman, Amawa, I. M., Sa'dijah, C., & Anwar, L. (2025). Developing adaptive digital book (ADB) in enhancing students' numeracy literacy ability. *Infinity Journal*, 14(3), 571–586. <https://doi.org/10.22460/infinity.v14i3.p571-586>
- Zheng, L., Niu, J., Zhong, L., & Gyasi, J. F. (2023). The effectiveness of artificial intelligence on learning achievement and learning perception: A meta-analysis. *Interactive Learning Environments*, 31(9), 5650–5664. <https://doi.org/10.1080/10494820.2021.2015693>