



Beyond Clickers: An AI-Powered Student Response System for Capturing Mathematical Reasoning in Indonesian Mathematics Classrooms

Fiki Rahmita^{1*}, Amanah Khairiyah²

^{1*}Faculty of Tarbiyah and Teacher Training, Universitas Islam Negeri Kiai Ageng
Muhammad Besari Ponorogo, Indonesia

²Ministry of Finance of the Republic of Indonesia.

*Corresponding Author. Email: fikirahmita@uinponorogo.ac.id

Abstract: This study aims to examine the design and pedagogical logic of the Si•Xiang•Hui AI Interactive Classroom, an AI-powered student response system developed at Zhejiang Normal University, China, and to explore its potential implications for Indonesian mathematics education. The study employed a qualitative descriptive-analytic case study design. The case involved a hands-on demonstration of the system during a seminar for education administrators from Southeast Asia held at Zhejiang Normal University in September 2026. The system developer served as the key informant, while 34 adult participants, including the two authors, participated as participant observers. Data were collected through participant observation, an audio recording of the session, the developer's presentation slides, the delegation's activity report, and relevant policy documents. The data were analyzed using reflexive thematic analysis with a deductive orientation informed by the ICAP framework, formative assessment, and knowledge building, with triangulation conducted across data sources. The findings indicate that the system extends beyond conventional clickers in four ways: it pairs answer choices with spoken explanations; enables every student to pose questions that can be addressed through human–AI collaboration; assigns higher scores to explanations based on their quality (up to ten points, compared with one point for an answer choice); and provides analytics intended to classify responses according to Bloom's Taxonomy and SOLO taxonomy levels. Two areas for further development were identified: enabling students to revise their recorded arguments and making teacher verification of AI-generated answers and scores an explicit step in the assessment process. The findings suggest that the pedagogical logic underlying the system, rather than the technology itself, can be adapted to Indonesian mathematics classrooms using low-cost tools, provided that teachers receive adequate preparation, scoring criteria incorporate mathematical correctness, and student data are appropriately protected.

Article History

Received: 05-08-2026

Revised: 25-09-2026

Accepted: 08-10-2026

Published: 20-10-2026

Key Words:

Artificial Intelligence;
Formative Assessment;
Mathematical Reasoning;
Student Response
Systems.

How to Cite: Rahmita, F., & Khairiyah, A. (2026). Beyond Clickers: An AI-Powered Student Response System for Capturing Mathematical Reasoning in Indonesian Mathematics Classrooms. *Jurnal Paedagogy*, 13(4), 1779-1791. <https://doi.org/10.33394/jp.v13i4.22758>



<https://doi.org/10.33394/jp.v12i3.16397>

This is an open-access article under the [CC-BY-SA License](#).



Introduction

Meaningful mathematics learning depends on two things: broad participation and reasoning that teachers can actually see. In many classrooms, a few confident students answer and ask questions while most of the class stays silent. Teachers are left with an incomplete picture of what each student understands, especially in large classes. This problem matters for Indonesia. In PISA 2022, Indonesia's mean mathematics score was 366, well below the OECD average of 472 (OECD, 2023), which points to a need to strengthen mathematical reasoning in everyday teaching.



In mathematics, a correct answer does not always reflect correct understanding. Many errors come from systematic misconceptions. A common example is applying whole-number rules to decimals, so that 1.45 is judged to be larger than 1.5 (Roell et al., 2017). Misconceptions like this can only be diagnosed when teachers know why a student chose an answer. Formative assessment relies on exactly this kind of evidence about student thinking, which teachers then use to adjust their instruction (Black & Wiliam, 2018). A meta-analysis by Rittle-Johnson et al. (2017) also found that prompting students to explain their reasoning (*self-explanation*) produces small to moderate immediate gains in both conceptual and procedural knowledge in mathematics, with larger effects when students are scaffolded toward high-quality explanations, although evidence from regular classrooms is still limited.

Student response systems, often called clickers, have long been used to widen participation. Each student selects an answer on a device, and the teacher sees the distribution of answers right away. In a systematic review, Wood and Shirazi (2020) identified six themes in students' experiences with these systems: engagement, interaction, anonymity, questioning, instant feedback, and the benefits and limitations of the technology. They concluded that outcomes depend on how teachers design questions and feedback. A more recent review found no consensus on whether such systems improve academic performance and noted that gamified quiz platforms such as Kahoot! dominate both practice and research (Serrada-Sotil et al., 2025). The same review judged the methodological quality of the reviewed studies to be weak, with limited control of variables and little use of validated instruments. The use with the strongest pedagogical grounding is *peer instruction*, in which students answer on their own, discuss and justify their answers with peers, and then answer again (Schell & Butler, 2018).

The ICAP framework helps explain what conventional clickers leave out. Chi et al. (2018) distinguish four modes of cognitive engagement (passive, active, constructive, and interactive), with learning gains increasing from one mode to the next. Choosing an answer is active engagement. Explaining the choice is constructive, because students generate ideas that go beyond what they were given. Building on ideas through dialogue is interactive. Clickers that record only answer choices usually stop at the active mode. They capture *what* students choose, but not *why*.

Generative artificial intelligence (AI) opens up new possibilities. Systematic reviews show that ChatGPT is the most widely studied AI tool in mathematics education, used mainly for personalized instruction and for immediate assessment and feedback (Opesemowo & Adewuyi, 2024). Awang et al. (2025) distinguish tools built for specific mathematical tasks, such as solving equations or visualizing geometry, from platforms that offer broader pedagogical support, such as adaptive systems and generative AI. Large language models (LLMs) bring both opportunities and risks for education (Kasneci et al., 2023; Wardat et al., 2023). In mathematics, ChatGPT still struggles with complex problems in topics such as spatial geometry and derivatives (Almarashdi et al., 2024). For this reason, it cannot replace the pedagogical role of teachers and works best in hybrid models with structured teacher guidance (Turmuzi et al., 2026). In Indonesia, Sutrisno AB et al. (2025) found that many of the 13 prospective mathematics teachers they studied could not yet analyze ChatGPT's answers in depth, which suggests how much depends on teachers' ability to check AI output.

Research on AI in education has also tended to pay little attention to the pedagogical perspective of educators (Zawacki-Richter et al., 2019). Järvelä et al. (2023) argue that AI should be treated as a collaborative partner that strengthens human learning rather than replacing the people involved in it. In classrooms, this implies designing for teachers–AI



complementarity, in which AI extends what teachers can notice while judgment stays with them (Holstein et al., 2019). China is a particularly interesting context because the development of AI in education there is driven systematically by the state (Knox, 2020). This direction has been reinforced by the *Master Plan on Building China into a Leading Country in Education (2024–2035)* and the *AI + Education Action Plan (2026)*, which aims for deep integration of AI and education by 2030 (Zhang, 2026). One practical example is the Si·Xiang·Hui AI Interactive Classroom developed at Zhejiang Normal University (ZJNU). The system follows a “Teacher–AI–Student” collaborative model in which every student can choose an answer, explain it orally, raise a hand, and ask questions, while all of these interactions are recorded as process-assessment data. This system was selected as the case for three reasons. First, unlike answer-based response systems such as Kahoot! and chatbots used by individual students, it combines answer choices, spoken explanations, student questions answered with LLM support, and learning analytics in a single whole-class routine. Second, it was designed by education researchers in response to classroom problems rather than as a technology product, so its pedagogical model is explicit and open to analysis. Third, it is reportedly in regular use from primary school to university rather than being a laboratory prototype (Zhang, 2026).

Previous research leaves three questions unanswered. First, research on student response systems has examined mostly answer-based quizzes and their effects on participation and achievement (Wood & Shirazi, 2020; Serrada-Sotil et al., 2025), but it has not explained how such systems can capture the reasoning behind students' answers, which is precisely the evidence that formative assessment in mathematics requires. Second, research on AI in mathematics education has focused mostly on chatbots used by individual students (Awang et al., 2025; Opesemowo & Adewuyi, 2024), so little is known about how LLMs can support whole-class interaction in which every student's reasoning is recorded. Third, although AI is increasingly framed as a partner of teachers (Holstein et al., 2019; Järvelä et al., 2023), few studies have analyzed, at the level of design, how a classroom system divides judgment between the teacher and AI when assessing students' reasoning. The relevance of such systems for diagnosing mathematical reasoning in Indonesian classrooms has received even less attention.

Accordingly, this study aims to examine the design and pedagogical logic of the Si·Xiang·Hui AI Interactive Classroom through the lenses of ICAP, formative assessment, and knowledge building, and considers its potential implications for capturing students' mathematical reasoning in Indonesian classrooms. The study contributes in two ways. Theoretically, it offers a design-level analysis of how an AI-powered student response system can move from recording answers to recording reasoning, and it identifies the design tensions that matter most for mathematics: the monologic nature of recorded explanations, AI scoring without a correctness criterion, and the loss of anonymity. Practically, it translates the system's pedagogical logic into design propositions for Indonesian mathematics classrooms that can be implemented with low-cost tools and tested in future classroom research.

Research Method

This study employed a qualitative, descriptive-analytic case study design (Yin, 2018). The case was the Si·Xiang·Hui AI Interactive Classroom (思·享·慧 AI 互动课堂), an AI-powered student response system developed by Zhang Lixin and his team at the College of Education, Zhejiang Normal University. A case study was chosen because the aim was to understand the design and pedagogical logic of one system as it was designed and



demonstrated, not to test its effect on learning outcomes. It is a revelatory case in Yin's (2018) sense: the seminar provided rare first-hand access to a system that, to the authors' knowledge, has not been described in the English-language literature.

The study took place at Zhejiang Normal University in Jinhua, Zhejiang Province, China, during the *Seminar for Education Administrators from Southeast Asia*, organized by the Ministry of Commerce of the People's Republic of China and ZJNU from 10 to 23 September 2026. The system was demonstrated in the session *Optimization and Innovation of Classroom Teaching Empowered by AI Technology* on 20 September 2026. The developer ran the session as a hands-on workshop rather than a lecture, so the seminar participants took the role of students in a class using the system. The participants were 34 adult education professionals attending the seminar, including both authors, who attended as members of the Indonesian delegation. The key informant was the developer, who presented the design of the system, operated it as the teacher, and answered participants' questions. The authors acted as participant observers: like the other participants, they used the device to choose answers, record spoken explanations, and ask questions, while noting how each feature worked, how the teacher used it, and how participants responded.

Data were collected from five sources: (1) the authors' participant observation notes on using the device; (2) the developer's presentation slides (29 slides); (3) the Indonesian delegation's activity report, which describes the session; (4) Chinese policy documents on education and AI cited in the presentation; and (5) an audio recording of the first hour of the session (up to the break), made by the second author, and its automatically generated transcript.

Before coding, the data were selected using explicit criteria. Data were included if they (a) described a feature of the device, the teaching software, or the analytics platform; (b) showed how a feature was used in the session or how the developer explained its pedagogical purpose; or (c) reported the scoring rules or class records generated by the system. Data were excluded if they (a) concerned other seminar sessions or general program activities; (b) consisted of contributions by other participants, which were neither quoted nor attributed; (c) concerned the content of the demonstration questions rather than how the system handled the responses; or (d) were automatically transcribed passages that could not be verified against the audio. Policy documents were used only to describe the context in which the system was developed, not as evidence of its features.

The session was a training activity for adult professionals rather than a study designed by the authors, and the authors took part as ordinary participants without intervening in the teaching. The recording was made with the presenter's knowledge. Because the sample class record in the slides showed students' names and identification numbers, only anonymized figures from it are reported.

The data were analyzed using reflexive thematic analysis (Braun & Clarke, 2019) with a largely deductive, theory-driven orientation, in three stages. First, the features of the device and the platform were identified and coded. Second, each feature was mapped to its pedagogical function using the ICAP framework (Chi et al., 2018), formative assessment (Black & Wiliam, 2018), and *knowledge building* (Chen & Hong, 2016). The three frameworks answer different questions about a response: what cognitive activity it asks of the student (ICAP), what the teacher can do with it (formative assessment), and what the class can do with it (knowledge building). Third, the relationship between each function and mathematics learning was examined, and opportunities for further development were identified; these results, together with the literature and the Indonesian context, were used to



formulate the conditions for adapting the system's design logic. Trustworthiness was supported by triangulating data sources (observation, the session recording, slides, and the activity report) (Carter et al., 2014; Yin, 2018), by confirmation with the developer, and by researcher reflexivity (Lincoln & Guba, 1985). Because the authors took part in a program hosted by the institution that developed the system, their position could shape their views. Direct observations were therefore kept separate from the developer's reports: each feature was coded as observed (O) when the authors used or saw it in the session and as reported (R) when it appeared only in the slides or the developer's account (Table 1).

The scope of the data also sets limits on the findings. The data came from a demonstration in a seminar with adult participants, not from actual mathematics classrooms; the recording covered only the first hour of the session; the analytics platform was presented on slides rather than used; and the main sources of information were the developer and the host institution. The findings are therefore an analysis of design and pedagogical potential, not evidence of effectiveness.

Results and Discussion

The findings are presented in two parts. The first five subsections describe how the system elicits and records students' reasoning; the last five discuss the conditions for adapting its design logic to Indonesian mathematics classrooms.

Starting Point: Classroom Problems, Not Technology

The developer did not start from technology. The starting point was seven classroom problems identified in the presentation (Zhang, 2026): low student engagement; teacher-centered instruction; teachers' inability to follow each student's learning dynamically, comprehensively, and accurately; the small number of students who ask questions, both because few get the chance and because many are unwilling, afraid, or unable to do so; the difficulty of monitoring every conversation during group work; limited opportunities for students to express ideas orally; and the difficulty of recording each student's performance, even though process assessment should in principle make up at least 50% of a course grade. The developer added that smaller classes and blended learning had not solved these problems (Zhang, 2026). The design principles were to build on regular classroom teaching, respond to real teaching problems and needs, and combine current pedagogical ideas with appropriate AI technology. This starting point matters, because the effectiveness of student response systems depends more on pedagogical design than on the technology itself (Wood & Shirazi, 2020).

The Device and System Architecture

Each student holds a handheld device (应答器, *AI Responder*) labeled "Think & Share." It has a screen that shows the student's name and the instruction issued by the teacher, A/B/C/D answer keys, an UP key for raising a hand, and a microphone key that is pressed and held while speaking. Responses are sent by radio frequency (RF) to a receiver (*AI Host*), passed via USB to the interactive teaching software on the teacher's computer, and then sent over the internet to a cloud server and a cloud-based LLM service. Class interaction data are then processed on an analytics platform (Figure 1).

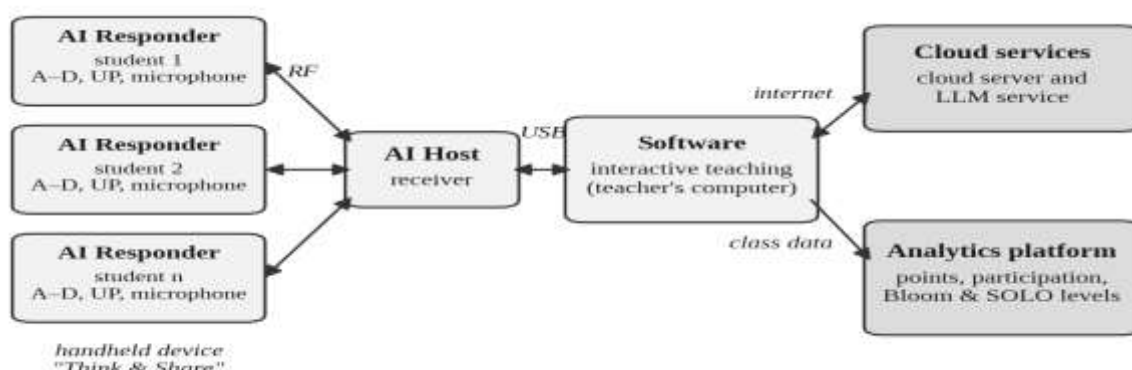


Figure 1. Architecture of the Si·Xiang·Hui AI Interactive Classroom
(adapted from Zhang, 2026)

This architecture is worth noting for schools with limited resources. The classroom does not need a phone or laptop for every student; one simple handheld device per student, a receiver, and one internet-connected computer for the teacher are enough. This is consistent with the developer's claim that the system was designed to be easy to operate, cost-effective, and widely applicable (Zhang, 2026). He described the devices as very cheap, although no price was given, and reported that the system is already used in primary and secondary schools and universities, with ten classrooms at ZJNU equipped with it.

The authors' experience as participants was broadly consistent with this claim. After a few minutes of explanation, participants could respond with only occasional reminders, although a few spoke when only a key press was required, which suggests that younger students would need a short practice routine. Each task followed the same routine: the teacher posed a question, a signal opened the response window, and students pressed a key or held the microphone key to speak. The class screen showed how many participants had not yet responded, so waiting (Zhang, 2026).

From Choosing to Explaining

Table 1. System features, pedagogical functions, and ICAP modes of engagement

Feature	Activity enabled	Pedagogical function	ICAP mode
A/B/C/D keys (O)	All students choose an answer at the same time	Equal participation; the class-wide answer distribution becomes visible	Active
Microphone: "explain your choice" (O)	Students record the reason for their choice	Makes thinking visible; helps diagnose reasoning	Constructive
Microphone: open expression (O)	Answering open questions, describing, reading aloud	Develops the ability to express ideas	Constructive
UP key and spoken questions (O)	Every student can raise a hand and ask a question	Builds a culture of questioning	Constructive
Human-AI collaborative answering (O)	The teacher selects students' questions and has the LLM answer them	Responds quickly to many questions	Interactive
AI summary of spoken responses (O)	The LLM condenses all responses into a few main points, which the teacher displays and discusses	Makes the range of ideas in the class visible; a basis for knowledge building	Interactive, if discussed
Recording of group discussion (R)	Each group's ideas are collected and shared	<i>Knowledge building</i> : from group ideas to individual understanding	Interactive
Point records (O) and analytics platform (R)	All interactions are stored and scored for each student	Process assessment and progress monitoring	—



Source: compiled from Zhang (2026), the session recording, and the authors' observations. O = observed in the session; R = reported in the slides or by the developer only.

Table 1 shows how the features of the system map onto pedagogical functions and ICAP modes of engagement. The main finding is that the system shifts the focus of a student response system from the *answer* to the *reasoning*. In the demonstration, each multiple-choice question was followed by the prompt “explain your choice.” For example, participants were asked about their attitude toward university students using AI to write papers and then explained their reasons orally (Zhang, 2026). This “choose, then explain” pattern moves students from the active to the constructive mode, and recording group discussions opens the way to the interactive mode (Chi et al., 2018).

This pattern is consistent with evidence that prompting students to explain their reasoning improves mathematics learning (Rittle-Johnson et al., 2017) and with the logic of *peer instruction*, which asks students to justify their answers (Schell & Butler, 2018). The difference is that every student's explanation is recorded and available to the teacher, rather than heard only by the person sitting next to them. The trade-off is that explaining to a device is a monologue rather than a dialogue: unless the recorded explanations are brought back into class discussion, the system secures constructive engagement but not necessarily interactive engagement (Chi et al., 2018). For teachers, these explanations are much richer formative assessment data than an answer distribution (Black & Wiliam, 2018). For student questions, the developer applies the principle of “everyone asks, questions are shared, and answers come from human–AI collaboration.” Questions are judged on clarity, depth, how thought-provoking they are, and originality, while spoken expression is judged on clarity, fluency, logic, and originality. Neither set of criteria refers to correctness, an issue discussed in the subsection on human–AI collaboration.

Process Assessment That Rewards Reasoning

The system records every interaction and assigns it points (Table 2). Every response earns one participation point, even when a student presses the microphone key and says nothing. Beyond that, points depend on the type and quality of the response: explaining a choice or expressing an idea can earn up to ten points, compared with one point for choosing an answer. In the session, the LLM assigned these quality points automatically, the teacher could add bonus points for particularly good answers, and responses could be ranked by score so that the strongest appeared first (Zhang, 2026). The weighting thus rewards the quality of reasoning, not merely the act of responding.

Table 2. Point system for classroom activities

Recorded activity	Points per response
Each participation in an activity	1
Choosing an answer	1
Reading-aloud evaluation	up to 4
Asking a question	up to 5
Explaining a choice	up to 10
Spoken expression	up to 10
Bonus points	set by the teacher

Source: Zhang (2026). Apart from the participation point, the values are maximum; actual points depend on the quality of the response.

A sample record shown by the developer lists 22 students from a university class (possibly a partial list), with total scores of 112–180 per student, 23–40 activities, 7–12 answer choices, and 13–21 spoken contributions each, but only 0–5 questions (Zhang, 2026).



Three points stand out. First, process assessment, which is hard to record by hand, is documented for every student. Second, the points track quality rather than quantity. Expression points ranged from 26 to 78 (a median of about 3.5 points per spoken contribution) and showed no positive association with the number of contributions (Spearman's $\rho = -0.23$, $n = 22$): the student with the most activities (40) and 20 spoken contributions earned the fewest expression points (26), whereas a student with only 13 contributions earned 70. This coefficient is weak and not statistically significant ($p \approx .30$), so it indicates the absence of a positive relationship rather than a genuine negative one. In practical terms, speaking more often did not by itself earn a student more expression points: the scores appear to depend on the quality of each contribution rather than on the number of contributions, which is what a reasoning-oriented scoring scheme intends. Because the record comes from a single class and the scoring prompt was not disclosed, this pattern describes the record rather than validating the AI scoring. Third, the number of spoken contributions mainly reflects how often the teacher asked for them, since every student responds to each prompt, whereas questions depend on students' own initiative. Asking questions remained the least frequent activity, even though the device made it easy.

This last observation fits the review by Wood and Shirazi (2020), which treats questioning as a separate theme in students' experiences with response systems. Technology can lower the practical barriers to asking questions, but a culture of questioning still has to be built through lesson design, for instance by setting aside time for questions, modeling good questions, and rewarding thoughtful ones. The question-quality criteria used at ZJNU could serve as a starting rubric for this purpose.

Learning Analytics at the Level of Cognition

Interaction data are processed on an analytics platform that reports activity points, variety of interaction, learning engagement, instructional constructiveness, levels of Bloom's cognitive objectives, SOLO levels, level of spoken expression, and an analysis and summary of students' spoken responses (Zhang, 2026). The slides showed these only as menu items and tabs, without examples or an account of how levels are assigned, so they are treated here as reported features whose validity has yet to be examined. If these classifications prove reliable, class data would go beyond frequency counts and point toward the cognitive quality of student responses. Mapping responses to the SOLO taxonomy (Biggs & Collis, 1982) is useful in mathematics because it separates responses that mention a single aspect (unistructural) from those that connect several concepts into a coherent whole (relational). From a *knowledge-building* perspective, data of this kind allow students' ideas to be treated as objects that can be compared and improved together (Chen & Hong, 2016). The summarizing function, by contrast, was used repeatedly: the LLM condensed participants' views on a good teacher into a few points, which the developer compared with his Chinese students' views as a form of shared knowledge building (Zhang, 2026).

The validity of these classifications is the central question, and the developer did not explain how it is established. Classifying a response as relational on the SOLO taxonomy, or as constructive in ICAP terms, requires judging the structure of the reasoning, not its length or fluency. In the decimal example discussed in the next subsection, a response such as “0.3 has fewer digits” relies on a single cue and is unistructural, whereas “0.3 is three tenths, or thirty hundredths, which is more than twenty-five hundredths” connects place value, equivalence, and comparison into a coherent argument and is relational. An LLM can make this distinction reliably only if it is given explicit level descriptors and anchored examples for the mathematical task, and only if its labels are checked against independent judgments by



trained teachers. Research on automated scoring treats such human-machine agreement, typically reported as Cohen's kappa or a similar index, as basic evidence of validity (Zhai et al., 2020). Without this evidence, an AI label of “relational” or “constructive” may reflect surface features such as length, vocabulary, or fluency rather than the quality of reasoning. For mathematics classrooms, SOLO and ICAP labels generated by AI should therefore be treated as suggestions for the teacher to confirm, not as assessment results.

Relevance for Mathematics Learning

The value of the system is clearest in mathematics. Imagine a teacher asking, “Which is larger, 0.25 or 0.3?” with the options A (0.25), B (0.3), and C (they are equal). Students who choose A probably hold the “longer is larger” misconception, applying whole-number rules to decimals (Roell et al., 2017). Yet students who choose B do not necessarily understand place value either. Some may reason that 0.3 is larger because it has fewer digits after the decimal point, a rule that would also lead them to judge 0.3 larger than 0.35. Only the students' explanations can separate these two groups. With AI summaries of those explanations, a teacher can group the class by reasoning rather than by right and wrong answers, and then plan targeted follow-up. The logic resembles a two-tier diagnostic test that pairs answer choices with reasons, except that here the reasons are open-ended and given in the students' own words.

Human-AI Collaboration and the Role of the Teacher

Human-AI collaborative answering addresses a real constraint. A single teacher cannot respond to every question in a large class and may face questions outside their expertise. In the session, the teacher acted as a gatekeeper, choosing the clearest student questions to send to the LLM; when a vague question produced an answer unrelated to ZJNU, he used it to show that AI answers are only as good as the questions put to them (Zhang, 2026). No step for checking an answer before it was shown to the class was observed. For mathematics content, however, this feature needs careful design. As noted in the Introduction, LLMs still make errors on complex mathematical problems (Almarashdi et al., 2024; Wardat et al., 2023), and prospective mathematics teachers may not yet be able to check AI answers in depth (Sutrisno AB et al., 2025). A Teacher-AI-Student model in which the teacher stays in charge is in line with calls for hybrid models (Turmuzi et al., 2026) and with the view of AI as a collaborative partner (Järvelä et al., 2023). In mathematics, the teacher's role in checking AI answers should be made explicit, for example through a step in which the teacher reviews an AI answer before it is shared with the class. The same applies to AI-assigned points: because they count toward grades and no scoring criterion refers to correctness, a fluent but mathematically wrong explanation could score well. A mathematics version would need a correctness criterion and teacher moderation of AI scores, and its fluency criterion should not penalize students explaining mathematics in a second language. Recordings of students' voices are also personal data, so their use should follow ethical principles for AI in education such as privacy, transparency, and accountability (Nguyen et al., 2023).

Room for Development: Recording Changes in Reasoning

While using the device, the authors noticed one opportunity for development that matters for mathematics. In the configuration used in the session, once a spoken explanation had been recorded, it could not be extended or recorded again. In *peer instruction*, however, some of the most important learning happens when students revise their answers after discussion (Schell & Butler, 2018). *Knowledge building* likewise treats ideas as objects to be continually improved (Chen & Hong, 2016). A “revise my argument” feature, or a follow-up



recording, would let teachers see not only students' initial reasoning but also how that reasoning changes. In mathematics education, this trajectory of change is valuable evidence of learning.

Counterpoints and Conditions for Transfer

Three counterarguments temper this analysis. First, recording, scoring, and ranking every utterance turns classroom talk into data, and critics of learning analytics warn that such systems can entrench inequalities and leave students little control over data about them (Selwyn, 2019). Second, points and rankings motivate unevenly. The developer presented them as a way to push reluctant university students to respond (Zhang, 2026); gamification improves performance on average, yet some students experience it as a source of anxiety and jealousy (Bai et al., 2020), and students may learn to give the fluent, point-earning answer rather than their actual reasoning. Because every response is linked to a named student, the system also forgoes anonymity, one of the themes in students' experience of response systems (Wood & Shirazi, 2020). Third, the system grew out of a state-led drive for AI in Chinese education (Knox, 2020) and out of local concerns, so it cannot simply be transplanted. These counterarguments set conditions for transfer rather than cancelling the system's logic: low stakes, transparency about what is recorded and why, and adaptation rather than wholesale adoption.

Lessons for Mathematics Classrooms in Indonesia

The main lessons from this case lie in its pedagogical logic rather than in the device. They are offered as design propositions to be tested rather than as prescriptions. First, make “explain your choice” a habit: any multiple-choice question in a mathematics class can be followed by a request for reasons, spoken or written. Second, design distractors around common misconceptions so that the class's answer distribution is diagnostic from the start. Third, give explanations more weight than answer choices in process assessment, judge them for correctness as well as clarity, and keep the stakes low. Fourth, start with tools that are already available, such as online forms, interactive quiz apps, and voice recorders on phones, and use LLMs to summarize students' reasons under teacher verification. Speech recognition should first be tested in Bahasa Indonesia and with mathematical language, and written reasons may work better for symbolic work. Fifth, prepare teachers, not just tools: designing diagnostic questions, reading data on reasoning, and checking AI output should be part of pre-service mathematics teacher education and professional development. Sixth, be transparent with students about what is recorded and why, and protect student data in line with Law Number 27 of 2022 on Personal Data Protection, which requires parental consent for processing children's personal data (Republic of Indonesia, 2022).

Conclusion

The Si·Xiang·Hui AI Interactive Classroom captures students' reasoning, not only their answers. It does so through four design features: each answer choice is followed by a spoken explanation, every student can ask questions that are answered through human–AI collaboration, explanations are weighted up to ten times more than answer choices in process assessment, and an analytics platform is intended to classify responses by cognitive level. These features move whole-class responses from the active to the constructive mode of engagement, but they reach the interactive mode only when recorded explanations are brought back into class discussion. For mathematics, the system has three gaps: recorded explanations cannot be revised, teacher verification of AI answers and scores is not an explicit step, and the scoring criteria do not include mathematical correctness. In this human–



AI collaboration, the teacher remains the decision-maker: the teacher chooses which questions go to the AI, judges whether AI answers and scores are mathematically sound, and decides how recorded reasoning is used to adjust instruction, while AI extends what the teacher can notice. For Indonesian mathematics classrooms, what can be transferred is the pedagogical logic of the system, namely “choose, then explain” and low-stakes process assessment that rewards reasoning, which can be implemented with low-cost tools without the device itself.

Recommendation

Future research should test adaptations of this model in Indonesian mathematics classrooms through *design-based research*, focusing on the quality of students' mathematical reasoning, the culture of questioning, and teacher workload. Priority should be given to checking the agreement between AI-generated scores and SOLO or ICAP labels and teachers' judgments, and to comparing named and anonymous responses in terms of students' willingness to speak.

Developers of AI-powered student response systems, including Si·Xiang·Hui, should consider a feature that lets students revise recorded arguments, an explicit step in which the teacher verifies AI answers and scores before they are shared or counted, a mathematical correctness criterion in the scoring rubric, and transparent documentation of how cognitive levels are assigned. Mathematics teachers can apply the propositions in the previous section immediately, starting with the habit of asking for reasons after every multiple-choice question and reviewing any AI-generated answer or score before it is shared with students or counted toward grades.

Acknowledgment

The authors thank the Ministry of Commerce of the People's Republic of China (MOFCOM) and Zhejiang Normal University for organizing the program, Prof. Zhang Lixin for sharing the presentation materials, the Ministry of State Secretariat of the Republic of Indonesia for selecting and sending the delegation, and UIN Kiai Ageng Muhammad Besari Ponorogo and the Ministry of Finance of the Republic of Indonesia for their support.

References

- Almarashdi, H. S., Jarrah, A. M., Abu Khurma, O., & Gningue, S. M. (2024). Unveiling the potential: A systematic review of ChatGPT in transforming mathematics teaching and learning. *Eurasia Journal of Mathematics, Science and Technology Education*, 20(12), em2555. <https://doi.org/10.29333/ejmste/15739>
- Awang, L. A., Yusop, F. D., & Danaee, M. (2025). Current practices and future direction of artificial intelligence in mathematics education: A systematic review. *International Electronic Journal of Mathematics Education*, 20(2), em0823. <https://doi.org/10.29333/iejme/16006>
- Bai, S., Hew, K. F., & Huang, B. (2020). Does gamification improve student learning outcome? Evidence from a meta-analysis and synthesis of qualitative data in educational contexts. *Educational Research Review*, 30, 100322. <https://doi.org/10.1016/j.edurev.2020.100322>
- Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy (Structure of the Observed Learning Outcome)*. Academic Press.



- Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice*, 25(6), 551–575. <https://doi.org/10.1080/0969594X.2018.1441807>
- Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, 11(4), 589–597. <https://doi.org/10.1080/2159676X.2019.1628806>
- Carter, N., Bryant-Lukosius, D., DiCenso, A., Blythe, J., & Neville, A. J. (2014). The use of triangulation in qualitative research. *Oncology Nursing Forum*, 41(5), 545–547. <https://doi.org/10.1188/14.ONF.545-547>
- Chen, B., & Hong, H.-Y. (2016). Schools as knowledge-building organizations: Thirty years of design research. *Educational Psychologist*, 51(2), 266–288. <https://doi.org/10.1080/00461520.2016.1175306>
- Chi, M. T. H., Adams, J., Bogusch, E. B., Bruchok, C., Kang, S., Lancaster, M., Levy, R., Li, N., McElloodon, K. L., Stump, G. S., Wylie, R., Xu, D., & Yaghmourian, D. L. (2018). Translating the ICAP theory of cognitive engagement into practice. *Cognitive Science*, 42(6), 1777–1832. <https://doi.org/10.1111/cogs.12626>
- Holstein, K., McLaren, B. M., & Aleven, V. (2019). Co-designing a real-time classroom orchestration tool to support teacher–AI complementarity. *Journal of Learning Analytics*, 6(2), 27–52. <https://doi.org/10.18608/jla.2019.62.3>
- Järvelä, S., Nguyen, A., & Hadwin, A. F. (2023). Human and artificial intelligence collaboration for socially shared regulation in learning. *British Journal of Educational Technology*, 54(5), 1057–1076. <https://doi.org/10.1111/bjet.13325>
- 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, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Knox, J. (2020). Artificial intelligence and education in China. *Learning, Media and Technology*, 45(3), 298–311. <https://doi.org/10.1080/17439884.2020.1754236>
- Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. SAGE Publications.
- Nguyen, A., Ngo, H. N., Hong, Y., Dang, B., & Nguyen, B.-P. T. (2023). Ethical principles for artificial intelligence in education. *Education and Information Technologies*, 28(4), 4221–4241. <https://doi.org/10.1007/s10639-022-11316-w>
- OECD. (2023). *PISA 2022 results (Volume I): The state of learning and equity in education*. OECD Publishing. <https://doi.org/10.1787/53f23881-en>
- Opesemowo, O. A. G., & Adewuyi, H. O. (2024). A systematic review of artificial intelligence in mathematics education: The emergence of 4IR. *Eurasia Journal of Mathematics, Science and Technology Education*, 20(7), em2478. <https://doi.org/10.29333/ejmste/14762>
- Republic of Indonesia. (2022). *Undang-Undang Republik Indonesia Nomor 27 Tahun 2022 tentang Pelindungan Data Pribadi* [Law of the Republic of Indonesia Number 27 of 2022 on Personal Data Protection]. <https://peraturan.bpk.go.id/Details/229798>
- Rittle-Johnson, B., Loehr, A. M., & Durkin, K. (2017). Promoting self-explanation to improve mathematics learning: A meta-analysis and instructional design principles. *ZDM*, 49(4), 599–611. <https://doi.org/10.1007/s11858-017-0834-z>



- Roell, M., Viarouge, A., Houdé, O., & Borst, G. (2017). Inhibitory control and decimal number comparison in school-aged children. *PLOS ONE*, 12(11), e0188276. <https://doi.org/10.1371/journal.pone.0188276>
- Schell, J. A., & Butler, A. C. (2018). Insights from the science of learning can inform evidence-based implementation of peer instruction. *Frontiers in Education*, 3, 33. <https://doi.org/10.3389/educ.2018.00033>
- Selwyn, N. (2019). What's the problem with learning analytics? *Journal of Learning Analytics*, 6(3), 11–19. <https://doi.org/10.18608/jla.2019.63.3>
- Serrada-Sotil, J., Huertas Martínez, J. A., & Granado-Peinado, M. (2025). Do audience response systems truly enhance learning and motivation in higher education? A systematic review. *Humanities and Social Sciences Communications*, 12(1), 1767. <https://doi.org/10.1057/s41599-025-06042-w>
- Sutrisno AB, J., Pratama, E. Y., Putra R, A., Nasution, S. H., & Lestyanto, L. M. (2025). Students' mathematical justification abilities in analyzing ChatGPT's answers. *Infinity Journal*, 14(2), 445–460. <https://doi.org/10.22460/infinity.v14i2.p445-460>
- Turmuzi, M., Azmi, S., & Kertiyan, N. M. I. (2026). ChatGPT in school mathematics education: A systematic review of opportunities, challenges, and pedagogical implications. *Teaching and Teacher Education*, 170, 105286. <https://doi.org/10.1016/j.tate.2025.105286>
- Wardat, Y., Tashtoush, M. A., AlAli, R., & Jarrah, A. M. (2023). ChatGPT: A revolutionary tool for teaching and learning mathematics. *Eurasia Journal of Mathematics, Science and Technology Education*, 19(7), em2286. <https://doi.org/10.29333/ejmste/13272>
- Wood, R., & Shirazi, S. (2020). A systematic review of audience response systems for teaching and learning in higher education: The student experience. *Computers & Education*, 153, 103896. <https://doi.org/10.1016/j.compedu.2020.103896>
- Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). SAGE Publications.
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1), 39. <https://doi.org/10.1186/s41239-019-0171-0>
- Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: A systematic review. *Studies in Science Education*, 56(1), 111–151. <https://doi.org/10.1080/03057267.2020.1735757>
- Zhang, L. (2026, September 20). *AI-empowered optimization and innovation in classroom teaching: A new model of “Teacher–AI–Student” human–machine collaborative teaching and learning* [Conference presentation]. Seminar for Education Administrators from Southeast Asia, Zhejiang Normal University, Jinhua, China.