NET WORK LISER

Discussion Paper Series

IZA DP No. 19025

October 2026

Making AI Tutoring Work for Poorer Schools

Hai-Anh H. Dang
World Bank, IZA@LISER, GLO,
Indiana University, LSE and UEH

Minh N. N. Do
Vietnam National University

The IZA Discussion Paper Series (ISSN: 2365-9793) ("Series") is the primary platform for disseminating research produced within the framework of the IZA@LISER Network, an unincorporated international network of labour economists coordinated by the Luxembourg Institute of Socio-Economic Research (LISER). The Series is operated by LISER, a Luxembourg public establishment (établissement public) registered with the Luxembourg Business Registers under number J57, with its registered office at 11, Porte des Sciences, 4366 Esch-sur-Alzette, Grand Duchy of Luxembourg.

Any opinions expressed in this Series are solely those of the author(s). LISER accepts no responsibility or liability for the content of the contributions published herein. LISER adheres to the European Code of Conduct for Research Integrity. Contributions published in this Series present preliminary work intended to foster academic debate. They may be revised, are not definitive, and should be cited accordingly. Copyright remains with the author(s) unless otherwise indicated.

IZA DP No. 19025 October 2026

Making AI Tutoring Work for Poorer Schools*

Abstract

Generative AI can make schoolwork look better without making students better at it. AI tutors could reach families who cannot afford human tutors, but whether they narrow learning gaps depends on policy. Randomized trials in Ghana, Nigeria and Sierra Leone show that AI-supported programs can raise learning measured without AI. Yet they do not isolate what works, and a large trial in rural China found no gain from AI personalization over non-personalized practice. This essay suggests that education systems, in Africa and beyond, judge complete programs rather than chatbots: by what students learn without AI, who benefits, and how they compare with feasible alternatives at full cost.

JEL classification

I21, I24, I25, O15, O33

Keywords

AI tutoring, generative AI, learning, inequality, poorer countries, Africa

Corresponding author

Hai-Anh H. Dang
hdang@worldbank.org

* We would like to thank Gero Carletto, Gabriel Demombynes, Jed Friedman, Peter Lanjouw, and various colleagues at and outside the World Bank for helpful discussions. We are grateful to the UK Foreign, Commonwealth and Development Office (FCDO) for funding assistance through the Data and Evidence for Tackling Extreme Poverty (DEEP) Research Program.

From chatbots to programs

Generative AI can make schoolwork look better without making students better at it. In a field experiment in Türkiye, about 1,000 high-school students practiced mathematics with a standard GPT-4 chatbot, with a tutor version prompted to give hints, or with no AI[1]. Practice grades rose by 48% with the chatbot and by 127% with the tutor. Yet on the exam that followed, taken without AI, students who had practiced with the chatbot scored 17% below those who had practiced without it. Those with the tutor showed no detectable exam advantage over students without AI. Fütterer *et al.* have argued that AI should coach students' self-regulated learning rather than supply answers[2]. We take that case as given and ask a different question: what do complete AI tutoring programs deliver in poorer schools, for whom, and at what cost?

The question matters because tutoring works but is scarce. Across 89 randomized trials of tutoring for children from preschool to grade 12, learning rose by 0.29 standard deviations (SD) on average. Effects were larger when teachers or paraprofessionals tutored and when tutoring took place during the school day[3]. Yet human tutors are costly and unevenly spread. Where schools fall short, families buy private tutoring. But children in poor households and poor places get the least[4].

AI tutors change these economics. They are available at any hour. They can run on low-cost phones and cost little, about US$5 per student for a WhatsApp mathematics tutor in Ghana[5]. In principle, they could bring tutoring to families that could never pay for it and narrow learning gaps. They also have clear weaknesses. They can hand over answers instead of prompting effort, and they make errors. They lack a human tutor's relationship, encouragement and accountability, and they need devices, power and connectivity that the poorest schools often lack. Whether AI tutoring narrows or widens gaps will therefore depend on policy: who gets access, how use is guided and which students programs target.

Most causal evidence on AI in education comes from richer countries, often from brief trials of students working one-on-one with a chatbot. A Stanford review of more than 800 papers available by October 2025 found only 20 with strong causal evidence. All but two came from high-income countries[6]. Randomized trials in Ghana, Nigeria and Sierra Leone now add program-level evidence from poorer countries. They show that AI-supported programs can improve assessed learning. The next task is to establish which programs add value, for whom, and at what full cost, not simply to expand access to chatbots.

What recent West African trials show

Three randomized trials compared AI-supported programs with usual schooling, using tests taken without AI (Table 1). In Ghana, students in grades 3 to 8 used Rori, a mathematics tutor on WhatsApp, in two 30-minute study-hall sessions a week for eight months, and their mathematics scores rose by 0.36 SD[5]. In Benin City, Nigeria, first-year senior secondary students worked in pairs with Microsoft Copilot in up to two 90-minute after-school sessions a week for six weeks. A composite score of English, AI knowledge and digital skills rose by 0.31 SD[7]. In Sierra Leone, students in grades 7 and 8 worked in pairs with Gemini in half of their mathematics lessons for eight weeks, and mathematics scores rose by 0.26 SD[8].

The programs shared some features: scheduled sessions, an adult in the room and devices at school. But the adult's role ranged from fixing technical problems in Ghana to leading the lesson in Sierra Leone, where teachers set objectives and starter prompts and closed with a class discussion. The tutors were built or framed to prompt students' own work. Rori gives a hint before a worked solution, and 76% of Gemini's messages in Sierra Leone were questions, against 2% direct solutions. Nigeria used a general-purpose chatbot with teachers' prompts on the English curriculum.

Some caveats are in order. Under each study's assumptions, instrumental-variable estimates imply a gain of about 0.03 SD for each additional day attended in Nigeria and of 0.38 SD for Sierra Leone students who completed the planned 12 hours. Neither is a general return to another session. These are effects of particular packages against usual schooling, not of AI alone or of any single feature. The Ghana and Nigeria trials carry a high risk of bias, and the tool's developer was involved in the Ghana and Sierra Leone studies[9].

Easy explanations do not fit

It is tempting to credit the structure: set times, supervision and tutors that hint rather than tell. A large trial in rural China cautions against that reading (Table 1)[10]. In 315 classes in 105 primary schools in low-income counties of Sichuan province, students in grades 4 to 6 practiced mathematics on school computers for 11 weeks. They had two 40-minute sessions a week, typically supervised by their mathematics teachers. One group used an AI system that tracked each student's knowledge, chose problems to match it, and offered hints and step-by-step guidance. Another used non-personalized practice software on the same schedule, and a third continued its usual activities. The AI version did no better than either (−0.04 SD against both), and the estimates rule out all but trivial gains.

The trial ran in 2021, before generative AI reached classrooms, so it cannot test conversational tutors. It does show that set times, supervision and hints were not enough there, and that AI personalization added nothing over non-personalized practice. Exposure and comparison conditions matter too. In China, both software groups had the same scheduled sessions, but recorded use was low and unequal (Table 1). Nigeria added after-school sessions; Ghana redirected existing study-hall time to mathematics practice. Extra subject practice remains a plausible contributor, but neither trial isolates it from AI.

Nor is unrestricted AI always harmful: US undergraduates who drafted an essay with off-the-shelf AI, rather than with search alone, tended to score higher on a test without AI a week later[11]. No general design rule has yet been established. One conjecture deserves testing rather than assuming: interaction, with a conversational tutor or with a teacher, may matter more than adaptive sequencing of problems alone.

How large, and how certain?

A recent review pooled eight comparisons of AI tutoring with human support or guardrails, four of them from poorer countries[9]. On tests meant to be taken without AI, the gain was 0.21 SD (95% confidence interval 0.05 to 0.36; Fig. 1). The evidence is positive, but the headline should not rest on the statistical significance of the full pool. The estimate is 0.16 SD (−0.04 to 0.37) without the two trials at high risk of bias and 0.17 SD (0.00 to 0.33) without the two whose testing conditions are unclear. Imprecision is not evidence of no effect.

Outcomes range from mathematics, physics and English tests to Nigeria's composite, and effects vary widely across settings and comparison conditions ($I^2 = 76\%$), so the pool cannot identify which design feature works. Time horizons also differ: the pooled trials test students at the end of a session, program, or school year, and none measures learning a year later.

Another review[12], prepared for the World Bank's 2026 World Development Report, covered 14 randomized trials of adaptive and AI-enabled tools in ten economies, none of them low-income. It found an average gain of 0.125 SD and no evidence that generative tools outperform the adaptive technologies that preceded them. Earlier technology offers a benchmark. Before generative AI, Mindspark raised mathematics scores in Delhi by 0.37 SD in 4.5 months, combining adaptive software with group instruction by teaching assistants in after-school

centers[13]. AI tutoring must demonstrate value against feasible alternatives, considering learning, full cost and who is reached.

Testing who benefits

In two of the three West African trials, students with higher baseline achievement gained more. In Sierra Leone, the effect rose by 0.195 SD for each standard deviation of baseline score; in Nigeria, higher achievers and students from better-off homes gained more[7,8]. Nigerian girls also gained more than boys, a result possibly driven by one girls-only school. In rural China, the AI version may have done worse than non-personalized practice for middle and higher achievers[10]. Without deliberate design, AI tutoring could reproduce the pattern of private tutoring, which favors those already ahead.

Equity has three layers: access to devices, power, and connectivity; participation; and learning. The trials report mainly the third, averaged over all students offered a program, although internet and power disruptions were common in both Nigeria and Sierra Leone[7,8]. The evidence shows neither that shared school devices reach the poorest students nor that home use widens gaps; both are questions for evaluation. Evaluations could prespecify subgroup estimates with their uncertainty and record who cannot attend or use the system. Mindspark, which paired level-matched software with teaching assistants, shows what is possible: gains were similar across the achievement range, which meant large relative gains for the weakest students[13].

Judge programs, not chatbots

A license buys access; evaluation must establish learning. We suggest that ministries and funders treat AI tutoring as a program to be tested rather than a product to be bought. Staged pilots could compare each program with the best feasible alternative[9], such as structured pedagogy[14],

phone tutoring[15], human tutoring or existing adaptive software. Much of the evidence so far compares AI only with no AI[6]. Independent tests could be taken without AI and, as Fütterer *et al.* urge[2], separate immediate performance from later retention. Large commitments and claims of lasting benefits, unlike reversible pilots, would best await a durability check a year later. Where possible, trials could add an arm that delivers the same program without AI, as the Chinese trial did with non-personalized software. This separates the effect of the whole package from the value that AI adds.

Cost comparisons could include staff time, devices, connectivity, and maintenance, not only licenses. Ghana's US$5 per student is a marginal cost, and Nigeria's US$48 per student covered a six-week pilot[5,7]. Trials could randomly vary instructional features and measure students' attempts, explanations, and revisions to investigate how learning changes. Programs would also need student-data protection and checks of accuracy and accessibility in local languages; one student switching Gemini into Krio in Sierra Leone shows possibility, not quality[8].

Equity, too, can be designed in. To narrow gaps, programs could start where private tutoring is scarcest, supply devices and connectivity at school, and begin each student at their own level.

These suggestions are not specific to Africa. Middle-income systems from South Asia to Latin America face similar conditions: large learning gaps, few trained tutors, tight budgets, and rising student use of chatbots. The trials in Türkiye and China show that the risks travel, and Mindspark in India shows that well-designed technology-aided instruction can pay off. Wherever governments buy AI for schools, the same tests seem sensible: judge whole programs against realistic alternatives, by what students learn without AI and by who benefits.

West African classrooms suggest that AI-supported programs can raise learning. The next generation of trials can tell governments which programs are worth paying for, and for whom.

References

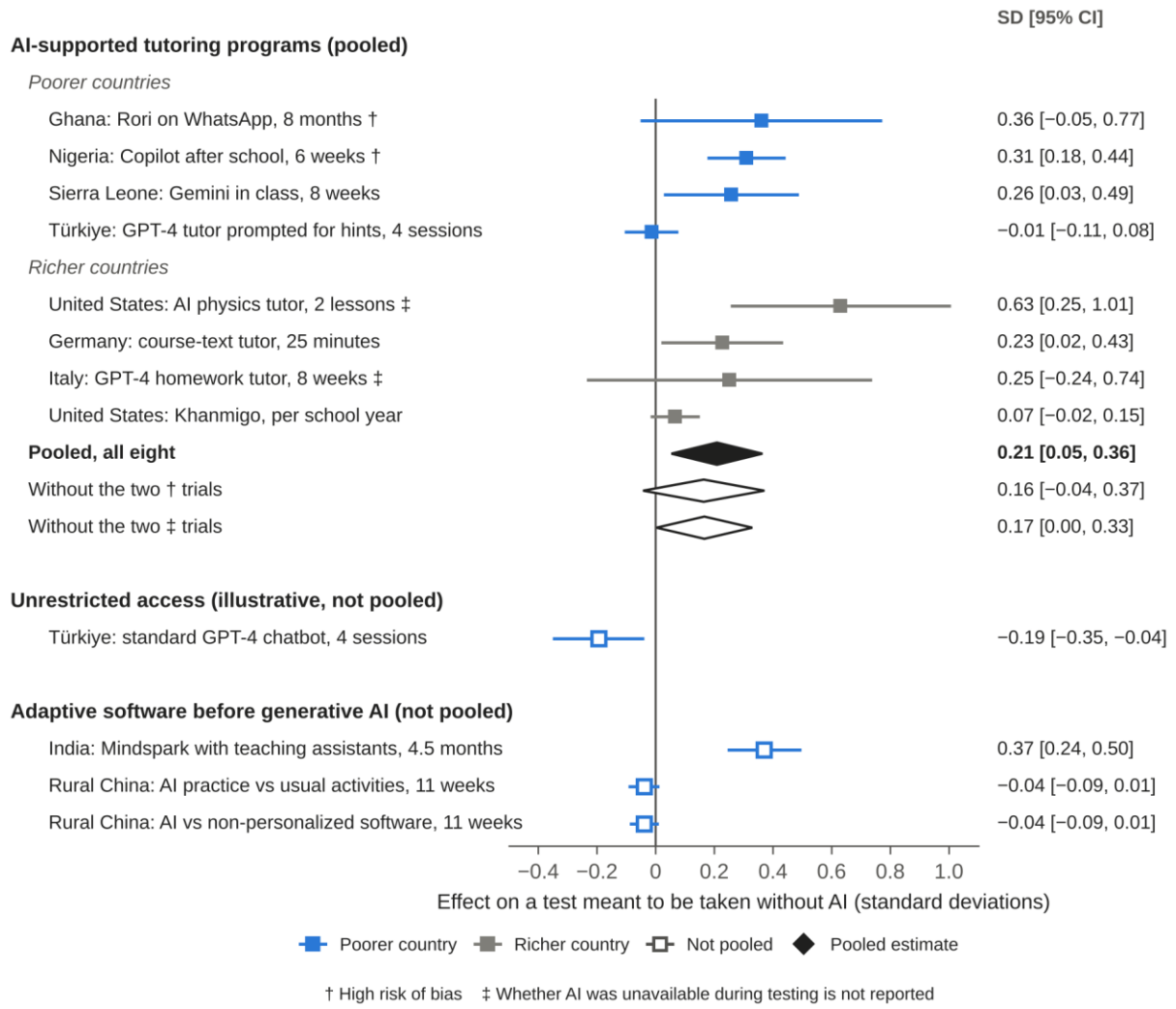
1. Bastani, H. *et al. Proc. Natl Acad. Sci. USA* **122**, e2422633122 (2025).
2. Fütterer, T. *et al. Nat. Hum. Behav.* https://doi.org/10.1038/s41562-026-02524-2 (2026).
3. Nickow, A., Oreopoulos, P. & Quan, V. *Am. Educ. Res. J.* **61**, 74–107 (2024).
4. Dang, H.-A. & Rogers, F. H. *World Bank Res. Obs.* **23**, 161–200 (2008).
5. Henkel, O., Horne-Robinson, H., Kozhakhmetova, N. & Lee, A. in *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky* (eds Olney, A. M. *et al.*) 373–381 (Springer, 2024).
6. Fesler, L., Martinez, J., Agnew, C. & Loeb, S. *The Evidence Base on AI in K-12: A 2026 Review*. Report (AI Hub for Education, SCALE Initiative, Stanford University, 2026); scale.stanford.edu/research-in-action/understanding-evidence-base-ai-k12-education.
7. De Simone, M. E. *et al. From Chalkboards to Chatbots: Evaluating the Impact of Generative AI on Learning Outcomes in Nigeria*. Policy Research Working Paper 11125 (World Bank, 2025).
8. LearnLM Team & Fab AI. *Teaching with Gemini: Measuring the Impact of Guided Learning on Student Mathematics Progress in Sierra Leone*. Technical report (Google DeepMind, 2026); storage.googleapis.com/deepmind-media/LearnLM/learnLM_sierraleone_may26.pdf.
9. Dang, H.-A. & Do, M. N. N. *Can AI Improve Household Welfare in Poorer Countries? Evidence and Measurement Challenges*. IZA Discussion Paper 18984 (IZA, 2026).
10. Ma, Y. *et al. Putting AI to the Test: Evidence from a Large-Scale RCT in Rural China*. CESifo Working Paper 12837 (CESifo, 2026).
11. Contractor, Z. & Reyes, G. *Experimental Evidence on the Learning Impact of Generative AI*. Preprint at arxiv.org/abs/2607.08849v2 (2026).
12. Burneo, A., Dinarte-Diaz, L., Lopez, C. & Molina, E. *Can EdTech Close Learning Gaps? Global Evidence from Digital Interventions*. Background paper for the World Development Report 2026 (World Bank, 2026); thedocs.worldbank.org/en/doc/2fba81cd6cd60d2f54532fc7062395fb-0050062026/original/Can-EdTech-Close-Learning-Gaps.pdf.
13. Muralidharan, K., Singh, A. & Ganimian, A. J. *Am. Econ. Rev.* **109**, 1426–1460 (2019).
14. Global Education Evidence Advisory Panel. *2023 Cost-Effective Approaches to Improve Global Learning* (FCDO, World Bank, UNICEF & USAID, 2023); documents1.worldbank.org/curated/en/099008110232520373/pdf/IDU-9401bfdd-f4ff-486b-84b4-570be4f65543.pdf.
15. Angrist, N., Bergman, P. & Matsheng, M. *Nat. Hum. Behav.* **6**, 941–950 (2022).

Table 1. Programs and comparison conditions in four randomized AI tutoring trials

No.	Feature	Ghana: Rori	Nigeria: Microsoft Copilot	Sierra Leone: Gemini	Rural China: adaptive software
1	Students and length	Grades 3–8; 11 schools; 8 months	First year of senior secondary; 9 schools; 6 weeks	Grades 7–8; 48 classrooms in 12 schools; 8 weeks	Grades 4–6; 315 classes in 105 schools; 11 weeks
2	Comparison group	Regular mathematics instruction	Usual schooling, without the after-school program	Usual mathematics lessons	Non-personalized practice software on the same schedule; usual activities (third arm)
3	Scheduled time	Two 30-minute sessions a week in existing study-hall periods; class hours unchanged	Up to two 90-minute after-school sessions a week	Two of four weekly mathematics periods (90 minutes)	Two 40-minute school-day sessions a week
4	Adult's role	Technical help; students work alone	Guides pairs with starter prompts; no direct teaching	Leads the lesson: objective, prompts, wrap-up	Usually the mathematics teacher supervises; no instruction
5	Tutor's response	A hint first, then a worked solution	General-purpose chatbot; teachers set the prompts	Questions in 76% of messages; solutions in 2%	Chooses problems to match estimated knowledge; hints and step-by-step guidance
6	Who gained more	No subgroup estimates	Higher achievers; better-off students; girls (tentative)	Students with higher baseline scores	No positive subgroup effect detected; possibly worse than non-personalized software for middle and top achievers (exploratory)
7	Cost per student reported	About US$5 (marginal only)	About US$48 for six weeks (US$9 marginal)	Not reported	Not reported

Sources: refs 5, 7, 8 and 10, in column order. *Note:* All four trials were randomized and tested learning without AI (effects are in Fig. 1). Only the Chinese trial compared the program with non-personalized software on the same schedule, and none followed students a year after the program. Recorded use in China averaged 262 minutes with AI and 183 minutes with non-personalized software over the 11 weeks, against 80 minutes a week scheduled.

Figure 1. Effects of selected AI-supported learning programs vary across settings and comparison conditions



Sources: ref. 9 (Tables A8 and A12), ref. 10 (Table 1) and ref. 13 (Table 2). *Note:* Squares are estimates with 95% confidence intervals, in standard deviations of tests meant to be taken without AI. Filled squares are pooled; hollow squares are shown for context and are not pooled. Diamonds are random-effects estimates from ref. 9, with small-sample (modified Hartung–Knapp) intervals. Three variances rest on assumptions (Ghana's clustering, a bound for US physics and a one-sided P value for Italy); varying Ghana's clustering gives 0.20 to 0.22. Poorer countries are low- and middle-income economies.