GENERAL INFORMATION

Title

Harnessing AI to Spark Curiosity: Experimental Evidence from Middle School Education

Principal Investigators

Vojtech Bartos (University of Milan), **Jana Cahlikova** (Erasmus University Rotterdam), **Miroslava Hapalova** (Scio Research), **Roman Lyach** (Scio Research), **Ipek Mumcu** (University of Exeter Business School)

Scio Research Team:

- **Matúš Kurian** co-conceptualized the theoretical framework of curiosity and contributed to the development of the Scio Research self-reported curiosity measure and the Curiosity-Oriented Inquiry Ability curiosity measure.
- **Hana Krulišová** co-conceptualized the AI-based curiosity-enhancing tool and contributed to its pedagogical design.
- **Lucie Vrbová** co-conceptualized the curiosity framework and contributed to the development of the Scio Research self-reported curiosity measure and the Curiosity-Oriented Inquiry Ability curiosity measure.
- **Helena Beranová** co-conceptualized the curiosity framework and contributed to the development and refinement of the Scio Research self-reported curiosity measure and the Curiosity-Oriented Inquiry Ability curiosity measure.

Country

Czech Republic, Slovakia

Region

All Czech Republics, all Slovakia

Start date November 24, 2025

End date July 1, 2026

Keywords

AI, curiosity, education

JEL Codes

C93, I21, D91

Abstract

In a clustered randomized field experiment, we study if a novel Generative AI (GenAI) tool accompanied by a methodological manual for teachers aimed at fostering student curiosity has a medium-term effect on curiosity and standardized test scores of students. The intervention is aimed at 8th grades of Czech and Slovak schools (N=124 schools, approximately 3,600 students who are aged 13-14). We collaborate with Scio Research, a subsidiary of a major Czech educational institution Scio, that develops the AI tool and carries out the randomized field experiment. To measure curiosity, we develop a novel behavioral measure building on psychological models of deprivation (D-) and interest (I-) type epistemic curiosity. We measure the effects of the intervention on student curiosity and on standardized test scores.

INTERVENTIONS

Intervention(s)

The intervention evaluates the impact of a Generative AI (GenAI)-based classroom tool designed to enhance curiosity of 8th-grade students from Czech and Slovak middle schools. The AI tool interacts with students through guided challenges and discussions that stimulate curiosity, creativity, and engagement with learning materials. The study tests whether structured exposure to curiosity-enhancing AI activities improves students' curiosity (primary outcome) and learning outcomes in standardized tests in mathematics and analytical thinking, while not affecting their mental health (secondary outcomes).

We employ a randomized controlled trial (RCT) with a staggered roll-out at the school level. Schools are randomly assigned to one of two groups:

1. **Early access (treatment)** – immediate access to the AI curiosity tool (February-March 2026). Teachers receive extensive training in AI tool use, receive accompanying methodological guidelines, and are constantly monitored by Scio

Research during the study period. A minimum of 15 minutes and a maximum of 60 minutes of in-class use weekly for the period of 3 months is contractually required and monitored.

2. **Delayed access (control)** – access to the AI curiosity tool after the endline data collection (September 2026).

Intervention (Hidden)

Intervention Start Date

2026-02-20

Intervention End Date

2026-06-30

PRIMARY OUTCOMES

Primary Outcomes (end points)

- Student's willingness to pay for D- and I-curiosity tasks in the behavioral measure of curiosity, measured in units of real-effort tasks, combined. We construct a simple z-normalized average over the six tasks (see behavioral tool description; as a general rule, all indices we construct follow the procedure as in Kling, Liebman, and Katz (2007) (1 variable). The outcome is measured only at the endline, in June 2026. If a student selects "I'm not interested" for a given task, WTP for that task is coded as zero. As a robustness check, for D-curiosity tasks where the follow-up response is that the student knows the answer to the question (i.e., not "I don't know the answer"), we assess sensitivity to coding that task as missing rather than zero.
- Behavioral measure of curiosity willingness to pay for D- and I-curiosity tasks separately. We construct a simple z-normalized average over the three tasks for each curiosity tasks, respectively (2 variables). The outcome is measured only at the endline, in June 2026.Behavioral measure of curiosity willingness to pay for D- and I-curiosity tasks separately. We construct a simple z-normalized average

over the three tasks for each curiosity tasks, respectively (2 variables). The outcome is measured only at the endline, in June 2026.

Primary Outcomes (explanation)

See description of the behavioral task in attachment. Measure builds on Alan and Mumcu (2024) measure of curiosity by eliciting willingness to pay in terms of real effort of students in a digital task. The tool distinguishes between D- and I-type curiosity (see secondary measures).

Multiple hypothesis testing (primary hypotheses). Since we test three primary hypotheses, we report both conventional ("per-comparison") p-values and p-values adjusted for multiple hypothesis testing. To address potential concerns about false discoveries arising from testing multiple outcomes, we apply the correction method proposed by Barsbai et al. (2020) using the authors' *mhtreg* Stata code. This approach extends the procedure developed by List, Shaikh, and Xu (2019) to allow for multiple-hypothesis correction in multivariate regression settings. The method explicitly accounts for the dependence structure among hypotheses (particularly relevant here as the combined index is by construction correlated with the two sub-components), thereby improving statistical power relative to traditional corrections that assume independence (e.g., Bonferroni, 1936; Holm, 1979).

SECONDARY OUTCOMES

Secondary Outcomes (end points)

- Student topic selection in the behavioral curiosity task. We test whether the treatment group exhibits shifts in topic interest towards particular subject areas relative to the control group.
- Newly developed behavioral measure of Curiosity-Oriented Inquiry Ability built on Abdelghani et al. (2022) and Lyach et al. (2025). Students will be invited to ask up to five questions after being presented a short text that stimulates curiosity. Each question will be evaluated using an AI tool for how much it reflects D- and I-type epistemic curiosity. Scores will be averaged across all submitted questions to produce two separate measures of student inquiry ability for D- and I- type curiosity. The measure will be validated using additional human evaluators and a

trained AI-coded evaluator in a small subsample of questions (two randomly selected questions from each student) submitted by control group students.

- Survey based measure of curiosity using Kashdan 5DCR measure (Kashdan et al. 2020). We construct an overall survey of curiosity as well as we distinguish between the six different domains of curiosity, following the original article.
- Newly developed survey-based method aimed at measuring curiosity of children based on two domains, D- and I-type curiosity, using eight survey items. See attached survey module.
- Test in mathematics and analytical skills. We use the overall standardized test score. The end-point test will be constructed similarly to the baseline test (attached).
- Mental well-being index using WHO-5 survey measure (Health Behaviour in School-aged Children study, 2023). We construct the index following the original article.
- Students' educational aspirations, include responses to two questions we utilize to construct academic pathway they aspire to continue. These questions include whether they would like to enroll in a more academic intensive secondary school track and whether they wish to enroll in university for post-secondary education.

Secondary Outcomes (explanation)

While the primary goal is to study the effect of the intervention on curiosity, we are also interested in understanding the effects on learning using standardized testing, student educational aspirations, and plans for university study. We are aware of possible adverse effects of increased Internet use on mental health (Braghieri et al., 2022 AER; Golin, 2022 Health Econ; Donati et al., 2025 J Health Econ). This is why we also measure student well-being. We have no priors on the effects as these could be both positive due to positive correlations between curiosity and learning. But they can as well be negative, where the AI tool usage crowds out other learning, leading to learning losses. Similarly, increased AI tool usage and increased time spent on screen (even given strict limits imposed and safety features implemented in the AI tool by design) could result in reduced well-being. A negative effect at any of the secondary outcomes (especially learning and well-being) would be informative of important trade-offs to be assessed. The randomized evaluation is a suitable tool to evaluate such trade-offs prior to a large-scale and permanent roll-out of the intervention.

Multiple hypothesis testing (secondary hypotheses). Alongside conventional ("per-comparison") p-values, we report p-values adjusted for multiple hypothesis testing to account for the increased likelihood of false discoveries when testing several secondary outcomes simultaneously. Specifically, we report two correction methods: (i)

Anderson's sharpened q-values, and (ii) the procedure of List, Shaikh, and Xu (2019), implemented through the *mhtreg* module of Barsbai et al. (2020), which extends the List-Shaikh-Xu approach to multivariate regression frameworks. The sharpened q-value procedure rescales the raw p-values to control the false discovery rate. This method offers a simple and flexible way to adjust for multiple testing even when individual regressions differ in their control variables, clustering structure, or other specification choices. Because the method requires only p-values as inputs, it is straightforward to implement and report. However, sharpened q-values do not account for the dependence structure among p-values, which can be important when outcomes are correlated. Hence, we also report results using Barsbai et al. (2020). This method accounts for dependence across hypotheses and therefore retains greater power than corrections assuming independence (e.g., Bonferroni, 1936; Holm, 1979). Researchers with a priori interest in specific outcomes may focus on the corresponding unadjusted p-values, whereas others should rely on the multiple-testing–adjusted results for a more conservative inference.

EXPERIMENTAL DESIGN

Experimental Design

Sample. 124 schools, 20 students per classroom on average in 180 classrooms (one, maximum two classrooms per school). Total of 3,600 8th grade students who are aged 13-14 years from Czech and Slovak middle schools ("zakladni skola" and "gymnazium"). We recruit schools via cold calls, emails, social media (Instagram, Facebook), and advertisements at major educational conferences and online media. The Slovak Ministry of Education actively supports recruitment. At the recruitment stage, we do not disclose the study's exact purpose but inform schools about the "free access to an AI tool and a methodology for developing soft skills", project duration, in-class testing, and surveys. All participating teachers will be remunerated for their participation, which among others helps reduce attrition.

Experimental manipulation. We employ a randomized controlled trial (RCT) with a staggered roll-out at the school level. Only schools that sign participation agreements, obtain parental consent and complete baseline data collection are included in the randomization pool and then are randomly assigned to one of two groups: **Early access (treatment)** – immediate access to the AI curiosity tool and **Delayed access (control)** – access to the AI curiosity tool after the endline data collection.

Timeline.

- May-November 2025: marketing campaign and school recruitment
- November 2025 – February 2026: registration of all participating schools, parental consents
- February 2026: Baseline data collection and randomization
- February/March 2026: Early access teacher training
- February/March-May 2026: intervention implementation by early access teachers with extensive monitoring by Scio Research
- June 2026: endline data collection
- September-November 2026: AI tool available also to control schools; without extensive monitoring by Scio Research

Outcomes. See above for both primary and secondary outcomes.

Other data collected. We collect data from multiple sources.

- *Country and regional level: an indicator for whether the school is in Czechia or Slovakia, and in which region (NUTS 3).*
- *School level:* open ministry databases on student numbers, 8th grade size, special needs, high achievers, foreign background, and location (linked to local SES). Pre-baseline.
- *Classroom level:* administrative records on participating class sizes and number of participating students (with parental consent). At baseline.
- *Teacher level:* AI use and knowledge. At baseline.
- *Student level* (on top of primary and secondary outcomes): student gender, student-reported parental education as a proxy for family SES, AI use and knowledge (in the endline, the question on AI tool usage will also include the option to select the AI tool used for the intervention), student aspirations (student plans for academic high school and university study). At baseline and endline.
- *AI tool level:* teacher logins and student logins and usage including number of messages sent, challenges completed, mind maps created, points earned, time

spent but not prompt/chat contents. Early access (treatment) group only, during the intervention period. While the system backend technically logs all interactions (including the full text of chats), the research team will not have access to this raw content.

Randomization balance. Following the concerns about balance test reporting (e.g., Bruhn and McKenzie, 2009), we use an omnibus joint test of orthogonality to test for balance using all baseline data described above. In a single OLS regression, we regress all the variables on the Early access (treatment) indicator. Then we test for joint significance of all the estimated coefficients using an F-test.

Manipulation checks. We measure student self-reported GenAI tool use and knowledge at the endline. We expect the usage (especially that for usage for school-related tasks) of the Early access group to be higher than that of the Delayed access group.

We expect that AI knowledge, number of types of AI tool use, and frequency of AI tools use in schools should be higher for the Early access group, relative to the Delayed access group.

Since we do not have login and usage data for the Delayed treatment group at the time of endline, we will at least examine correlations between Early access (treatment) group student logins and usage, and self-reported AI usage to assess the validity of self-reports using objective data.

We employ the regression analysis described below.

Standard Errors. Standard errors are clustered at the school level. We will further refine inference using small-cluster-sample corrections or resampling-based methods (CR2 corrections (Bell and McCaffrey, 2002; Cameron and Miller, 2015) or wild cluster bootstrap (Cameron et al., 2008).

Hypotheses. The primary null hypotheses are that:

- **Curiosity pooled.** Students' willingness to pay in the behavioral game (primary outcome, a measure of curiosity) in Early access group is statistically indistinguishable from the Delayed access group. The intervention is aimed at rejecting the null hypothesis.

- **D-curiosity.** Early access group students' willingness to pay for D-type of curiosity remain statistically indistinguishable between the two groups.
- **I-curiosity.** Early access group students' willingness to pay for I-type of curiosity remain statistically indistinguishable between the two groups.

Secondary hypotheses are that:

- **Self-reported curiosity measures.** The self-reported survey measures of curiosity remain statistically indistinguishable between the two groups.
- **Curiosity-Oriented Inquiry Ability.** The average score for D-type and I-type curious questions remain statistically indistinguishable between the two groups.
- **Learning outcomes.** That the standardized test scores in mathematics and analytical thinking remain statistically indistinguishable between the two groups.
- **Aspirations.** That the self-reported educational aspirations (plans for academic high school and university study) remain statistically indistinguishable between the two groups.
- **Well-being.** That well-being of students remains statistically indistinguishable between the two groups.

Validity of the behavioral measure. To investigate the validity of our behavioral measure, we will study correlations between our primary outcome (WTP index) and the self-reported curiosity indices (Kashdan et al. 2020; Scio Research, mimeo).

Further, the subscales of the Kashdan et al. (2020) and the Scio Research (2025) questionnaire measures are aimed at measuring the two curiosity types that our behavioral measure captures. So, we also study correlations between a WTP index for the D- and I-types of curiosity with the sub-index of the Kashdan et al. (2020) measure for Deprivation sensitivity and Joyous exploration, respectively. Using the behavioral measure for Curiosity-Oriented Inquiry Ability, we will also study the correlation between the degree of curious questions asked by the students and WTP index.

Moreover, we are interested in associations between our behavioral measure of curiosity and learning outcomes. We thus study correlations between the WTP index and test scores. We also study correlations with expressed aspirations.

All the correlations above are measured at the endline using the responses from the delayed access group (control group) only.

The willingness to pay measure also relies on the fact that the effort put in the real effort task is equal across groups. To test this, we measure response times for the real effort tasks as well as for the WTP choices. We report simple mean comparisons by treatment assignment for the two variables.

Regression analysis. We test the primary and secondary hypotheses using a linear regression using the student level-data from endline. We regress the outcome variables on the Early access (treatment) group indicator. We control for strata variables and all baseline measures of outcomes available (i.e. excluding the primary outcome that is only measured at the endline). In robustness checks, we also estimate the regressions (i) with strata variable controls only, (ii) without any additional controls, and (iii) with covariates selected using a post-double-selection LASSO. We make statistical inference based on results of a two-tailed test. Standard errors are clustered at the school level. The estimated effects are intent-to-treat effects (ITT).

Heterogeneity effects. While the sample size allows sufficiently powered tests on the full sample, we will still conduct exploratory heterogeneity analysis. We do this along the following dimensions:

- **Baseline teacher AI use, knowledge, and attitudes.** Median split of a z-score index of AI use and knowledge questions, with higher levels of knowledge and use of tasks increases the index. We expect that teachers with prior knowledge of AI will be better able to employ the tool. Hence, we expect larger effect sizes for classes of these teachers.
- **Baseline student AI use and knowledge.** Median split of a z-score index of AI use and knowledge questions, with higher levels of knowledge and use of tasks increases the index. We do not have a clear prior on the direction of the effect. On the one hand, higher baseline AI use and knowledge students may not experience additional benefits from the tool. On the other hand, prior experience with AI could be a necessary condition for effective AI tool usage.
- **Baseline teacher and student curiosity.** Median split of a z-score index of self-reported curiosity measure questions, with higher levels of curiosity increases the index. We expect that students and classes taught by teachers with lower levels of curiosity will benefit more from the tool. Hence, we expect larger effect sizes for these students and classes taught by these teachers.
- **Baseline ability.** Quintile split based on the baseline standardized test score.
- **Parental education.** Split by (i) both parents completed high school with a “maturita” state exam but no university degree, (ii) at least one parent completed university, or (iii) none of the above.

- **Student gender.** Alan and Mumcu (2024) document effects of educational intervention on curiosity being driven primarily by girls. We expect a similar pattern.

Note: All but the teacher-level heterogeneity splits rely on student-level data, while standard errors are clustered at the school level, so statistical power remains high for student-level outcomes.

Long-term effects measurement. Exploratory follow-up (not pre-registered confirmatory hypothesis). While the staggered roll-out of the AI tool does not allow us to measure clean experimentally controlled long-term effects going beyond the endline study, we may still be able to track the performance of students using the treatment assignment as an instrument. Since we expect the intensity of the usage of the tool to be higher in the Early access group (a measurable outcome using student and teacher logins) due to the intensive marketing and monitoring by Scio Research (see timeline above) and due to the fact that the AI tool is given to teacher who may not be teaching the same students in the following academic year, we can use treatment assignment as an exogenous source of variation in the AI tool usage. At a later stage, we aim collect data on student high school transmission. If available and if the identification strategy relying on treatment-difference in teachers' implementation intensity proves feasible, this result will be reported in a follow-up study.

Attrition. We will report student attrition rates at endline by treatment status and test for differential attrition regressing school drop-out on the Early access (treatment group indicator. If attrition differs, we will conduct Lee bounds (Lee, 2009) as robustness analysis.

We will also report student attrition by treatment interacted with balancing variables. We will report an F-test on the joint significance of the interaction terms between the treatment indicator and the balancing variables. This approach also allows us to comment on whether there is a change in composition of types of schools/students across our sample, even if the numbers of schools/students who attrit do not differ by treatment status.

Experimental Design Details [Header]

Randomization Method

Randomization. Software-based randomization (using Stata 19). Stratified randomization at the school level. Reproducible with a fixed seed (260209).

Stratification. Schools are randomly selected from a pool of schools that signed participation agreements, obtained parental consent and completed baseline data collection. We stratify randomization based on country (Czechia and Slovakia).

Randomization Unit

School level

Was the treatment clustered?

Yes, at school level

EXPERIMENT CHARACTERISTICS

Sample size: planned number of clusters

124 schools

Sample size: planned number of observations

3600 students in 124 schools

Sample size (or number of clusters) by treatment arms

62 schools (1800 students) in Early access (treatment) group

62 schools (1800 students) in Delayed access (control) group

Minimum detectable effect size for primary outcomes (accounting for sample design and clustering)

We can detect a minimum effect size of approximately 0.14 SD on the index of D- and I-curiosity between the early-access (treatment) and delayed-access (control) groups

with N = 3,600 students in 62 clusters per group, assuming α = 0.05, power = 0.80, and an intra-cluster correlation of 0.05. Using a more conservative intra-cluster correlation of 0.10 based on Alan and Mumcu (2024), the minimum detectable effect is estimated to be 0.18 SD.