The Returns to Tailoring: Evidence from AI in the Classroom

Jaime Polanco-Jiménez[1,2] Kristof De Witte[1,3]

[1]Leuven Economics of Education Research, KU Leuven, Belgium
[2]Pontificia Universidad Javeriana, Colombia
[3]UNU-Merit, Maastricht University, The Netherlands

July 30, 2025

Outline

Outline

This Paper in a Nutshell

Research Question

How does the **design of an AI tool** affect student learning? Specifically, what are the returns to deeply **tailoring** an AI chatbot to a specific curriculum versus using a **generic** AI tool?

Motivation

- Financial literacy is crucial for economic well-being, but knowledge gaps are wide and consequential (**Lusardi2014**).
- School systems face persistent teacher shortages and need effective, scalable educational technology (Pressley, 2021; Sutcher et al., 2019).
- The value of curriculum-tailoring for modern AI tools has not been causally tested.

How We Answer This

- Large-scale RCT with **N=2,440** Belgian secondary students.
- We compare three groups within classrooms:
  - **Control:** Traditional instruction.
  - **T1 (Generic AI):** Instruction + general-knowledge AI chatbot.
  - **T2 (Tailored AI):** Instruction + curriculum-specific adaptive AI chatbot.

A Preview of Key Findings

Tailoring is the critical ingredient for success.

Engagement:

- Tailored AI ↑ module completion by **10 pp**.
- Generic AI ↓ module completion by **5 pp**.

Durable Learning (LATE):

- For students *induced* to finish by the Tailored AI, knowledge retention two months later ↑ by **0.36 SD**.

Learning (ITT):

- Offering the Tailored AI robustly ↑ immediate test scores by **0.036 SD**.

Mechanisms:

- Tailored AI ↑ student self-confidence.
- Generic AI ↓ student self-confidence.

Our Contributions to the Literature

1. **First causal estimates on the returns to tailoring AI in education.**
   - By directly comparing a generic and a tailored tool, we show that deep curricular integration is a key determinant of effectiveness.
   - This provides a crucial insight as one-size-fits-all AI platforms (e.g., ChatGPT) become ubiquitous.

2. **Broadens the set of outcomes for evaluating EdTech.**
   - We document effects on *learning efficiency* and *long-term knowledge retention*, addressing the common "fade-out" problem of interventions.

3. **Provides evidence on the mechanisms driving effects.**
   - The tailored AI's success is mediated not just through instruction, but through its ability to foster student *engagement* (completion) and *self-confidence*.

Outline

Context: Financial Education in Flanders, Belgium

The Setting: A compelling environment to study educational interventions.

- **Mandated Curriculum:** A 2019 reform made financial literacy a key cross-curricular competence for all secondary students.
- **The Paradox:** Flemish students rank high on PISA financial literacy assessments on average, but this masks a severe achievement gap linked to socioeconomic status (SES).
  - Performance gap between high/low SES students is **104 score points** (vs. 87 OECD avg.).
  - A student's SES explains **16.8%** of the variance in financial literacy performance (vs. 11.6% OECD avg.) (**OECD2023**).
- **Implementation Challenges:**
  1. Taught by non-specialists, creating demand for high-quality, standardized resources.
  2. Significant student heterogeneity across academic tracks, requiring differentiation.

The Policy Tension

Can a single AI tool provide both **standardized quality** and **personalized instruction** to solve this?

Conceptual Framework and Related Literature

Educational Production Function

- Simply increasing instructional time often has small returns (Hanushek, 2003).
- The **quality and efficiency** of that time are the primary drivers of learning.
- Many educational interventions suffer from **effect fade-out** over time (**Lortie-Forgues2019**).

Educational Technology (EdTech)

- Evidence is mixed; replacing teachers with standard online formats can have null or negative impacts (Cacault et al., 2021; Figlio et al., 2013).
- A seminal study shows causally that ICT effectiveness depends critically on its **integration with the local curriculum** (**Goolsbee2006**).

AI in Education

- Meta-analyses confirm AI tools can have positive effects (Wang & Zhao, 2024).
- **Our key question:** Does the **Goolsbee2006** finding on curricular integration still hold? Is a generic AI sufficient, or is deep tailoring necessary to unlock its potential?

Outline

Experimental Design

- **What:** A large-scale Randomized Controlled Trial (RCT).
- **When:** January - May 2024.
- **Who:** **2,440** students in 120 classrooms across 58 secondary schools in Flanders, Belgium.
- **Topic:** Belgian personal income tax system.
- **Randomization:** At the **individual level within classrooms** to non-parametrically control for teacher, classroom, and peer effects.

- Control (T0): Traditional learning path using standard materials.
- Generic AI (T1): Condensed instruction + AI chatbot with general tax knowledge (not specific to Belgium).
- Tailored AI (T2): AI chatbot specifically designed for the Flemish curriculum and Belgian tax code, providing adaptive feedback.

Data Collection and Key Outcomes

- Data collected via online surveys at three points in time:
  - Pre-test (t=0)
  - Post-test (t=1, immediately after intervention)
  - Follow-up test (t=2, two months later)
- We construct three primary, standardized outcome variables:
  1. **Gained Financial Literacy:**

$$\text{Standardized (Post-Test Score - Pre-Test Score)}$$

  2. **Learning Efficiency:**

$$\text{Standardized} \left(\frac{\text{Gained Financial Literacy}}{\text{Time Spent on Module}} \right)$$

  3. **Knowledge Retention:**

$$\text{Standardized (Follow-up Test Score at t=2)}$$

- We also collect rich data on demographics, prior grades, and psychosocial scales (e.g., self-confidence, motivation).

A Key Finding: High and Differential Attrition

The central empirical challenge is that module completion is a key economic outcome that varies starkly across groups.

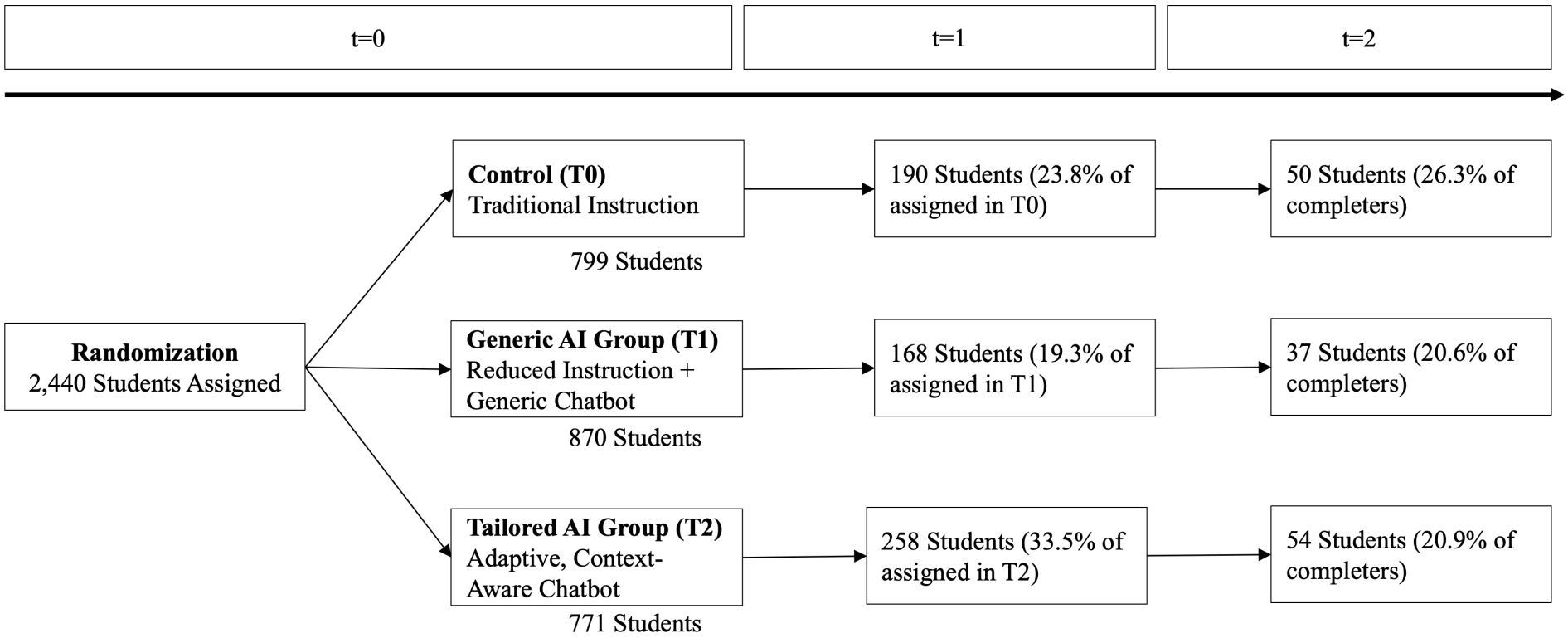

Figure: Experimental Design and Participant Flow

- **Tailored AI (T2) was most engaging: 39.0%** completion rate.
- **Generic AI (T1) actively disengaged students: 23.1%** completion rate.
- **Control (T0):** 29.4% completion rate.

A Key Finding: High and Differential Attrition

The central empirical challenge is that module completion is a key economic outcome that varies starkly across groups.

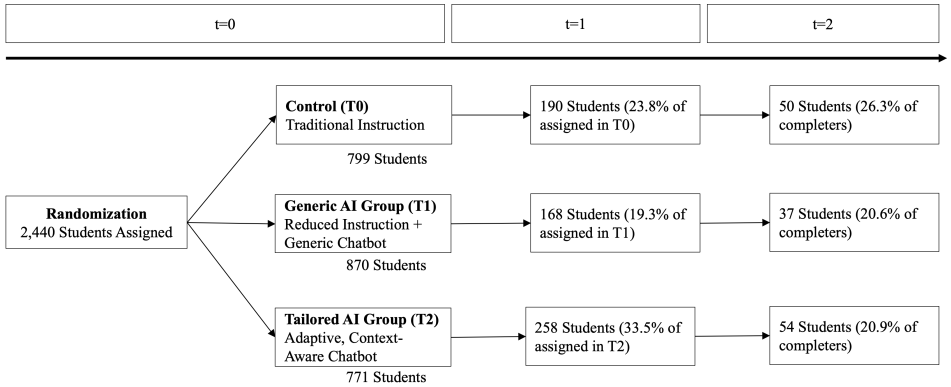

Figure: Experimental Design and Participant Flow

- **Tailored AI (T2) was most engaging: 39.0%** completion rate.
- **Generic AI (T1) actively disengaged students: 23.1%** completion rate.
- **Control (T0):** 29.4% completion rate.

Implication

Simple comparisons on the sample of completers would yield biased estimates. Our empirical strategy must account for this non-random selection.

Successful Randomization: Continuous Variables

We test for balance on the full randomized sample (N=2,440). The random assignment created statistically equivalent groups.

Variable	(1) Control Mean (SD)	(2) Generic AI Mean (SD)	(3) Tailored AI Mean (SD)	(4) p-val (T1-Ctrl)	(5) p-val (T2-Ctrl)
Pre-Intervention Outcomes					
Fin. Lit. Score (Pre)	0.349	0.337	0.343	0.451	0.723
	(0.242)	(0.246)	(0.238)		
Psychosocial Scales (1-5)					
Attitude/Motivation	2.866	2.829	2.796	0.315	0.108
	(0.733)	(0.735)	(0.749)		
Self-Confidence	2.683	2.686	2.693	0.932	0.814
	(0.867)	(0.837)	(0.866)		
Engagement/Commitment	2.548	2.491	2.507	0.127	0.301
	(0.776)	(0.711)	(0.779)		
Observations	799	870	771		

No statistically significant differences at baseline.

Successful Randomization: Categorical Variables

Balance also holds across all observable categorical characteristics.

Variable	Control (%)	Generic AI (%)	Tailored AI (%)	p-value (χ^2 test)
Gender				0.695
Female	50.45	48.41	47.00	
School Type				0.635
General (ASO)	68.03	65.82	64.27	
Technical (TSO)	29.30	31.65	32.00	
Last Math Grade				0.440
Over 70%	39.36	37.79	38.69	
Language at home				0.634
Dutch	83.43	83.27	81.10	

Conclusion

The randomization was successful. Any differences we find post-treatment can be attributed to the causal effects of the interventions.

Outline

Empirical Strategy: Addressing Attrition

Our primary challenge is the high and differential attrition. A naive comparison of completers is biased. We therefore use a multi-faceted approach.

1. **Intent-to-Treat (ITT) on the Full Sample (N=2,440)**
   - Measures the effect of being *offered* the tool. This is our main, policy-relevant estimate.
   - We estimate: $Y_i = \alpha + \delta_1 Z_{i,1} + \delta_2 Z_{i,2} + \mathbf{X}_i'\gamma + \mu_s + \eta_i$
   - To handle missing outcomes for attritors, our main specification makes a conservative assumption: **impute their knowledge gain as zero**. This provides a credible lower-bound estimate.

Empirical Strategy: Addressing Attrition

Our primary challenge is the high and differential attrition. A naive comparison of completers is biased. We therefore use a multi-faceted approach.

1. **Local Average Treatment Effect (LATE) via Instrumental Variables (IV)**
   - Measures the effect of *actually completing* the intervention for the subpopulation of "compliers".
   - We use the initial random assignment (Z_i) as an instrument for treatment completion (T_i).
   - This credibly estimates the Treatment-on-the-Treated effect without selection bias.

First Stage:
$$T_{is,k} = \pi_{0k} + \pi_{1k} Z_{is,1} + \pi_{2k} Z_{is,2} + \mathbf{X}is'\omega_k + \mu_s + \nu is, k \quad (1)$$

Second Stage:
$$Y_{is} = \beta_0 + \beta_1 \hat{T} is, 1 + \beta_2 \hat{T} is, 2 + \mathbf{X}is'\lambda + \mu_s + \eta is \quad (2)$$

Outline

First-Order Effect: AI Assignment Drives Student Engagement

The first stage of our IV analysis confirms that treatment assignment powerfully predicts module completion.

Table: The Effect of Treatment Assignment on Module Completion (First Stage)

	(1)	(2)
Dependent Variable:	Completed Post-Test (0/1)	
Assigned to Generic AI (T1)	-0.0504***	-0.0447**
	(0.0058)	(0.0211)
Assigned to Tailored AI (T2)	0.0973***	0.0968***
	(0.0119)	(0.0217)
Control Group Mean	29.4%	
Observations	2,440	2,440
R-squared	0.045	0.018
Baseline Controls	Yes	No
School Fixed Effects	Yes	No

Notes: OLS estimates on the full randomized sample. Robust standard errors, clustered by school, are in parentheses. *** p¡0.01, ** p¡0.05, * p¡0.1.

Key Takeaway

- Assignment to the **Tailored AI** increased the probability of completion by **9.7 percentage points**.
- Assignment to the **Generic AI** *decreased* the probability of completion by **5.0 percentage points**.

ITT: Modest but Robust Learning Gains from Tailored AI

Our main, conservative estimates use the full sample (N=2,440) and impute a gain of zero for all attritors. This provides a credible lower-bound for the population-level impact.

Table: Impact of AI Assignment on Learning Outcomes: ITT Estimates

	Full Sample (Zero-Imputed)			Completer Sample (for comparison)		
Dependent Var:	(1) Gain Score	(2) Learning Eff.	(3) Retention	(4) Gain Score	(5) Learning Eff.	(6) Retention
Generic AI (T1)	-0.0024	0.0333***	-0.2835	0.0625***	0.2584***	0.2685
	(0.0064)	(0.0102)	(0.4970)	(0.0178)	(0.0555)	(0.2214)
Tailored AI (T2)	**0.0356*** **	0.0260	-0.0256	0.0127***	0.0465**	**0.3872*** **
	(0.0062)	(0.0251)	(0.2640)	(0.0048)	(0.0182)	**(0.0643)**
Observations	2,440	2,440	2,440	616	616	141

Notes: All outcomes are standardized. All regressions include baseline controls and school fixed effects. Robust standard errors clustered by school.

Main Finding (Column 1)

Offering the **tailored AI tool** led to a statistically significant increase in the immediate knowledge gain score of **0.036 standard deviations** ($p < 0.01$) across the entire student population.

- **Potential Impact:** For those who complete the module, the tailored AI leads to a massive, durable learning gain of 0.387 SD in knowledge retention (Column 6).

Mechanisms: Tailored AI Boosts Self-Confidence

On the sample of completers, we find a striking divergence in how the AI tools affected student self-perception.

Table: Effect of AI Assignment on Psychosocial Outcomes (Completer Sample)

	(1) ITT on Completers
Panel A: Attitude & Motivation	
Assigned to Generic AI (T1)	-0.0889*
Assigned to Tailored AI (T2)	-0.0785***
Panel B: Self-Confidence & Self-Efficacy	
Assigned to Generic AI (T1)	**-0.1347*** **
	(0.0217)
Assigned to Tailored AI (T2)	**0.1976*** **
	(0.0413)
Observations	583-584

Notes: Dep. var. is the change in the psychosocial construct (1-5 scale). Full controls included.

A Key Mechanism

- The **Tailored AI** significantly ↑ student self-confidence by **0.20 points**. The supportive, adaptive nature of the tool seems to empower students.
- In sharp contrast, the **Generic AI** significantly ↓ self-confidence by **0.13 points**. A poorly contextualized tool can be actively harmful.

Effect for Compliers (LATE): Tailored AI Fosters Durable Learning

The LATE estimates show the effect for students whose completion behavior was changed by the intervention offer.

Table: The Effect of Treatment Completion on Learning Outcomes (LATE Estimates)

Dependent Variable:	(1) Gained Score (SD)	(2) Learning Eff. (SD)	(3) Retention (SD)
LATE Estimates (2SLS)			
Completed Generic AI (T1)	0.1824***	0.2118***	0.2819
	(0.0392)	(0.0241)	(0.2344)
Completed Tailored AI (T2)	0.0607*	0.0618**	**0.3583***
	(0.0275)	(0.0177)	**(0.0901)**
Observations	616	616	141
First-Stage F-statistic	> 1000	> 1000	> 1000

Notes: 2SLS estimates. Endogenous variables are completion indicators, instrumented with random assignment.

The Power of a Well-Designed Tool

For the students induced to complete the module by the **Tailored AI's** superior design, the intervention produced a significant and durable **0.36 standard deviation increase in knowledge retention** two months later.

Outline

Conclusion: The Returns to Tailoring are High

1. **Curricular integration is the critical ingredient for effective educational AI.**
   - A curriculum-tailored chatbot significantly boosted student engagement, while a generic version actively harmed it.
   - This initial behavioral response is the primary mechanism driving all subsequent learning outcomes.

2. **Well-designed AI has significant and durable learning effects.**
   - We find a robust, population-level (ITT) increase in immediate learning of **0.036 SD**.
   - For students on the margin of engagement (the compliers), the tailored AI had a much larger impact, increasing long-term knowledge retention by **0.36 SD**.

3. **Policy Takeaway: Be wary of "one-size-fits-all" AI.**
   - The promise of inexpensive, generic AI solutions may be illusory. To be effective, educational technology must first be used.
   - Our results suggest deep curricular integration is essential to solve the first-order challenge of student engagement and unlock the technology's pedagogical potential.

References I

Cacault, M. P., Hildebrand, C., Laurent-Lucchetti, J., & Pellizzari, M. (2021).Distance learning in higher education: Evidence from a randomized experiment. *Journal of the European Economic Association*, *19*(4), 2322–2372. https://doi.org/10.1093/jeea/jvaa060

Figlio, D., Rush, M., & Yin, L. (2013).Is it live or is it internet? experimental estimates of the effects of online instruction on student learning. *Journal of Labor Economics*, *31*(4), 763–784. https://doi.org/10.1086/669930

Hanushek, E. (2003).The failure of input-based schooling policies. *Economic Journal*, *113*(485), F64–F98. https://EconPapers.repec.org/RePEc:ecj:econjl:v:113:y:2003:i:485:p:f64-f98

Pressley, T. (2021).Factors contributing to teacher burnout during covid-19. *Educational Researcher*, *50*(5), 325–327. https://doi.org/10.3102/0013189X211004683

Sutcher, L., Darling-Hammond, L., & Carver-Thomas, D. (2019).Understanding teacher shortages: An analysis of teacher supply and demand in the united states. *Education Policy Analysis Archives*, *27*(35). https://doi.org/10.14507/epaa.27.3626

References II

Wang, L., & Zhao, M. (2024). *Can artificial intelligence technology promote the improvement of student learning outcomes?—meta analysis based on 50 experimental and quasi experimental studies* [Working Paper].

Thank You

Questions?

jaime.polancojimenez@kuleuven.be
kristof.dewitte@kuleuven.be