Journal of Development Economics
Registered Report Stage 1: Proposal

Impact Evaluation of the Darsel Math Personalized Learning Platform in Jordan

Authors:
Caroline Krafft (University of Minnesota, corresponding author, kraff004@umn.edu)
Paul Glewwe (University of Minnesota)
Tessa Bold (Institute for International Economic Studies, Stockholm University)
Abdulhamid Haidar (Darsel)
Ryan McWay (University of Minnesota)
Paola Giannattasio (Stockholm University & Institute for International Economic Studies)
Mojtaba Mahdipour (Department of Economics, Stockholm University)

Author order reflects time on project to date; it may be updated on stage 2 submission.

Date of Latest Draft: February 25, 2026

Abstract
Education contributes to economic growth, and it increases individuals' incomes and overall quality of life. Most developing countries have succeeded in enrolling almost all children in primary school. Yet in these countries student learning in primary school is often far below the levels envisioned for their grade. This challenge is also the case in Jordan; student performance in Jordan's primary schools remains well below international benchmarks. One strategy to address low student performance in developing countries is the use of "Educational Technology," often referred to as "EdTech," which can take many forms. One form that has attracted interest is the use of AI chatbots, which can personalize learning to teach at the right level (TaRL). We will implement a randomized controlled trial to assess whether a personalized math chatbot – Darsel – increases the mathematics and socio-emotional skills of grade 6 students in Jordan.

Keywords: Education Technology, Jordan, Teaching at the Right Level
JEL Codes: I25, O15, O33

Study pre-registration: Registered with AEA RCT Registry on November 21, 2025

Timeline:

- August 2025: Development of survey instruments with Ministry of Education (MoE), IRB approvals, MoE and research operations staff training
- September 2025: Start of school year: Data collection firm (proctor) training and baseline data collection, randomization and roll-out of intervention
- November 2025: AEA RCT registration
- Late November/Early December 2025: End of first semester: Midline - proctor refresher training and data collection
- April - May 2026: End of academic year: Endline - proctor refresher training and data collection
- June - 2026 - August 2027: Data cleaning, analysis, and working paper
- August 2027: Submission of the Stage 2 report
- September 2027: Preparation of policy brief(s), blog, dissemination

1 Introduction

Human capital formation contributes to economic growth and increases incomes, improves health, and raises overall quality of life at the individual level (Gethin 2025; Cutler and Lleras-Muney 2012). Education is a key component of human capital. Yet despite progress in educational enrollment, learning outcomes in developing countries remain low, with children performing well below grade level (Bold et al. 2017). Globally, an estimated 125 million children lack basic literacy and numeracy skills, even after four or more years of schooling (World Bank 2018). Described as a global "learning crisis," this situation has been exacerbated by COVID-19, especially for already disadvantaged students (Moscoviz and Evans 2022), and is now recognized as one of the most important policy challenges in education.

This learning crisis demands adaptive and remedial tools that can effectively boost basic learning. Even before COVID-19, ministries of education were adopting education technology (EdTech) with the goal of accelerating learning (Vegas et al. 2019). School closures during COVID-19 made the use of technology even more salient. Especially in low- and middle-income countries (LMICs), low-tech interventions emerged as an effective way to reduce learning losses and support education (Angrist et al. 2022). EdTech apps are a particular focus for marginalized groups, such as refugees (Koval-Saifi and Plass 2018b, 2018a). The evidence on the impact of EdTech overall, and in comparison to other interventions, is decidedly mixed, and contingent on other supports and inputs (Evans and Mendez Acosta 2021).

We will use a randomized controlled trial (RCT) to rigorously measure the impact of a promising EdTech intervention, Darsel, on mathematics learning, socio-emotional skills, and academic progress of students. We will test Darsel in grade 6 (ages 11-12) in 294 public schools across seven school districts in Jordan. Darsel has developed a personalized mathematics learning platform (AI chatbot) for students, even those with limited internet access, enabling them to practice mathematics through low-bandwidth channels such as WhatsApp. The platform promotes adaptive and remedial learning by providing individualized learning programs tailored to the learner, with the aim of bringing students' basic skills and abilities up to the level expected for their grade level.

Darsel is already at scale, used nationally in grade 7 in Jordan. In a pilot RCT for grade 6, Darsel increased math scores by 0.21 standard deviations. We will test a "business as usual" version of Darsel (T1), and an "enhanced" model (T2). T2 will provide additional support for teachers, as well as individualized (to the teacher, class, and student levels) pedagogical advice based on their students' successes and struggles with questions in the chatbot.

We will specifically investigate the following research questions:

1. What is the impact of an at-scale chatbot-based personalized and adaptive learning platform (Darsel) on middle-school math learning?

2. Can chatbots also improve student socio-emotional skills, such as motivation and self-efficacy?

3. Can chatbots be a cost-effective alternative to hardware-based EdTech interventions in LMICs?

4. How important is teacher support for EdTech adoption?

We will explore these questions and how the answers vary based on different student and teacher characteristics.

Jordan is an ideal context to test the potential of Darsel's model. Countries in the Middle East and North Africa, including Jordan, are some of the lowest performers in international assessments (El-Kogali and Krafft 2020; OECD 2023; von Davier et al. 2023). Despite high enrollment rates – close to 100% in grades 1 - 6 and 95% in grades 7 - 10, representing more than 1.8 million children in basic education (Department of Statistics (Jordan) 2023) – Jordan ranked 73rd out of 81 countries in the Programme for International Student Assessment (PISA) 2022 mathematics assessment (OECD 2023). In the 2023 Trends in International Mathematics and Science Study (TIMSS) math assessments, math scores for Jordanian 4th graders ranked 51st out of 58 participating countries, and 60% of students failed to reach the lowest international standards benchmark (von Davier et al. 2023).

Given their ready scalability, chatbots that leverage popular messaging platforms such as WhatsApp have enormous potential to improve learning outcomes. Past research on computer-assisted learning (CAL) and personalized instruction suggests that this direction has potential (Major et al. 2021; Rodriguez-Segura 2022; Zhang et al. 2020; Escueta et al. 2020), but rigorous evidence on the causal impact of chatbots on education is scant (Okonkwo and Ade-Ibijola 2021). Therefore, this study will make an important contribution to the research on how to improve learning and on the potential of EdTech, by providing some of the first rigorous evidence on the educational impact of chatbots.

Providing evidence on a program with national scale will also be a particularly valuable contribution, as Darsel is at national scale in Jordan in grade 7 and soon will be for our targeted grade 6 as well. A limitation of the existing CAL and EdTech literature is that evaluations are often of programs operated with high fidelity on a limited scale (Lai et al. 2015; Muralidharan et al. 2019; Yang et al. 2013; Mo et al. 2015; Lai et al. 2013). Even those evaluations at "large" scale are not for programs scaled up nationally (Bettinger et al. 2023; Naik et al. 2020). Assessing CAL and EdTech at a national scale is critically important for informing policy; education programs can have radically different implementation and effects at scale (Kerwin and Thornton 2021).

EdTech software, such as chatbots, is argued to be a very cost-effective intervention at scale (Muralidharan et al. 2019). Indeed, past research has identified CAL as a particularly cost-effective intervention at scale, compared to other interventions, especially since shifting from low to high-intensity personalization has no additional equipment costs (Major et al. 2021). As a point of comparison, only two out of fifteen interventions that were identified as effective in a recent review of educational interventions for LMICs had lower costs[1] than Darsel (McEwan 2015). The literature suggests that the introduction of at-home education technology, such as Darsel, provides a boost to in-classroom performance (Escueta et al. 2020; Soe et al. 2000; Verhoeven et al. 2020; Tzenios 2020).

There are other AI math learning chatbots, but only Darsel is focused on at-scale impact in LMICs by partnering with public schools. Rising Academies' chatbot "Rori" was built for a network of for-profit private schools (Henkel et al. 2024). It was tested using school-provided hardware (phones) in a one-hour-per week study hall for grades 3 - 9 in Ghana, and had an effect size of 0.36 for math learning (Henkel et al. 2024). Similarly, in Nigeria, an after-school program for upper secondary students used an AI chatbot in school computer labs and led to an effect size of 0.23 on (targeted) English skills (De Simone et al. 2025). Darsel is distinct in not requiring hardware or in-school facilitation from schools, removing major barriers to scale.

Darsel, by its personalized and remedial nature, is also a new model for "Teaching at the Right Level" (TaRL), which has been demonstrated to be highly effective through in-person modalities (Banerjee et al. 2016). A meta-analysis specifically of RCTs for personalized learning found the level of personalization to be critically important; programs with moderate levels of personalization had effect sizes of 0.13 SDs overall, compared to 0.35 SDs for personalization that included adapting or adjusting to learners' skills (Major et al. 2021). The evidence from our evaluation will also contribute to the debate on the effectiveness of EdTech programs as complementary tools or substitutes to traditional learning (Büchel et al. 2022; Cardim et al. 2023; Bettinger et al. 2023; Linden 2008; Beg et al. 2022). Furthermore, we will undertake comparisons of Darsel's cost-effectiveness to that of other EdTech interventions and, more generally, other types of education interventions for LMICs (e.g., Muralidharan et al. 2019; Kremer et al. 2013; Angrist et al. 2025). Due to the cost-effectiveness, the complementary educational boost of Darsel to traditional education, and the personalized TaRL model, we believe an evaluation of Darsel will provide important insights into the role of chatbot EdTech in combatting the learning crisis in LMICs.

[1] See Appendix C for a discussion of costs.

2 Research Design

2.1 Intervention

Darsel is an education technology nonprofit organization which is registered in California as a Public Benefit Corporation and is recognized by the U.S. Internal Revenue Service (IRS) as a 501(c)(3) tax-exempt organization. Its mission is to increase student learning in low-resource settings. It has developed a personalized math learning platform (chatbot) that allows students to practice and learn math using low-cost, low-bandwidth messaging channels (e.g., WhatsApp, Facebook Messenger). The platform is adaptive and remedial; Darsel's algorithms identify and work to rectify learning gaps. Because the Darsel model is already developed with personalized and dynamic targeting, in our study we cannot distinguish the benefits from (1) increased access to content (2) simple targeted instruction and (3) more dynamic, personalized instruction. AI is likely to be particularly valuable for (3) but its marginal value is uncertain and comparing these different channels is an important area for future research on chatbots.

The role of AI in the Darsel chatbot is limited to recommending the 'next question' based on a prediction of student mastery, which is based on previous response patterns. All content is selected from a pre-prepared dataset of programmatically-generated content. Unlike other chatbots, students are not conversing with a large language model (LLM), nor are they exposed to content that was immediately produced by an LLM. It is important to note that our research does not aim to measure the value-add of using AI, or make any claims about AI's contribution to any measurement of Darsel's impact. In this work we focus on the impact and cost-effectiveness of the existing Darsel model, as well as an enhanced model that provides additional teacher support.

Darsel has been implemented in public schools nationally in Jordan for grade 7. Students use Darsel from home, using a household device (e.g., cell phone), with no need to distribute new hardware and no disruption to school activities.[2] Math teachers receive usage reports for their students via WhatsApp and can access a web-based dashboard. The teacher's primary role is to motivate and encourage student usage of the Darsel platform.

Darsel has developed over 500,000 questions, hints, and explanations that are aligned with Jordan's national curriculum. The learning experience on Darsel revolves around questions and answers, where students receive curriculum-aligned questions and respond with the final answer. When students answer incorrectly, Darsel responds with hints (after the first attempt) and full explanations. Darsel also uses AI to dynamically select content for each student based on their

[2] Darsel is designed to be accessible even to the most vulnerable. As of 2016, 98% of students lived in a household with a mobile phone (authors' calculations based on 2016 Jordan Labor Market Panel Survey).

response patterns, with the objective of ensuring that content is always provided "at the right level," and in each student's proximal zone of development.

More specifically, the algorithm leverages expert-defined skill-related metadata to estimate a student's mastery probabilities for various skills. The personalization occurs continuously, so the student's learning path evolves with each question. All content is selected from a large database of content that has previously been reviewed and approved by a math expert. The role of LLMs is limited to content development and quality assurance. A/B testing is used to optimize the effectiveness of algorithms and content. The chatbot also offers motivational messages and gamification features to make the learning experience more fun and interactive. For example, students unlock weekly levels based on the number of correct questions and also get celebrated for streaks of correct questions.

Not only is Darsel's technology and content innovative but so is its method of delivery and implementation. The chatbot's simplicity, relying on popular messaging channels (WhatsApp, Facebook Messenger), enables it to be implemented in low-resource settings with common household devices and minimal teacher training. This makes it effective for students who are 'hard to reach' in traditional classroom settings, such as girls and refugees. Darsel does not disrupt school operations, nor is it demanding of teachers. Implementation of Darsel does not require teachers to adjust their lesson plans or change their approach to teaching. It only asks (but does not mandate or enforce) that teachers spend a total of five minutes per week to review the Darsel report and encourage students to use the platform. In Jordan, it has institutionalized gamification (e.g., school and district leaderboards) and incentives (award ceremonies) to maximize student engagement, motivation, and confidence.

Darsel's collaboration with Jordan's Ministry of Education has been ongoing since 2021, when the first pilot was conducted in two public schools. Darsel was then expanded gradually to national adoption for grade 7 students in over 2,000 government schools in March 2023. For grade 7 in the 2024-2025 school year, 43% of students (53,953 students) used Darsel at least once with proper registration.[3] Among registered students, nearly half used Darsel for five or more weeks and a quarter used Darsel across both semesters.

In this experiment, we will test the business-as-usual model of Darsel (T1) versus additional teacher support for teaching at the right level (T2). The second treatment arm (T2) includes a set of enhanced interventions designed to influence teacher behavior and improve instructional quality. Students' and teachers' access to the Darsel platform will be available during the entire academic year (September 2025 to May 2026) for all schools in the seven districts in the sample

[3] Darsel estimates another 26,605 students, 22%, used Darsel without proper registration; for grade 6 we will ensure all students have to register.

that are assigned to either T1 or T2. The reports and activities received by teachers will be based on the treatment to which they were randomly assigned.

2.1.1 Basic Teacher Support (T1)

Darsel's current model is designed to empower teachers to motivate at-home student usage of the Darsel platform. Teachers are invited to an optional virtual training session at the beginning of the school year. District officials also send packages to their schools which contain an 11-page teacher guide and a number of one-page student flyers (one per student) which provide brief information about Darsel as well as access instructions. Teachers are asked to review the teacher guide, explain Darsel to students, distribute the student guides, and ask students to take them home to their parents and start using Darsel.

Teachers are given access to school reports which they can generate and access through WhatsApp. These reports are focused on usage levels, including a leaderboard of students based on the number of weekly correct questions. Nudge messages are periodically sent to teachers to encourage them to generate a report. Teachers are also given access to a web-based platform which can be optionally used for additional reporting and administrative tasks. Furthermore, they are added to a WhatsApp group that includes a designated district-level math supervisor as well as all other math teachers who are participating in the same school district. Weekly district leaderboards, which rank schools based on average usage levels, are shared on this WhatsApp group. District supervisors are encouraged to motivate teachers in the group via WhatsApp, as well as through other channels, including calls and during their ordinary school visits (which are a major portion of supervisors' job responsibilities). Activities by district supervisors are not reported to Darsel, and neither district supervisors nor teachers are mandated by policy to any particular activity. Award ceremonies are conducted at the end of each term, with awards given to the top students, teachers, and districts, all based on usage rates.

Darsel's role during program implementation is limited to providing technical support and encouragement. Darsel's team will coordinate closely with the district-level supervisor and will also share messages on the district WhatsApp group to provide encouragement or to address technical questions. Darsel's team will also respond to individual messages from teachers but will not proactively visit schools or reach out to teachers in an individual capacity (outside the district-level group).

2.1.2 Advanced Teacher Support (T2)

Teachers in T2 will receive all the access and support provided to teachers in T1 and will also have access to additional product features and behaviors. Their reports will go beyond usage data and provide pedagogical insights on their students' learning levels and difficulties at the individual and classroom level. Their reports will also be accompanied by messages that summarize learning outcomes and provide specific pedagogical guidance and actionable recommendations. These messages constitute an automated information intervention aimed at encouraging evidence-based teaching practices, including teaching at the right level. Therefore, while reports for T1 are primarily designed to facilitate student usage, reports for T2 teachers also intend to affect teacher behavior and content prioritization during classroom hours.

In addition, teachers in T2 will participate in a single 1-hour in-person training session focused on promoting effective use of the data contained in the reports. This session will provide explicit instruction on how to interpret and apply the feedback to improve classroom engagement and learning. Finally, T2 teachers will receive proactive outreach and support through WhatsApp and phone calls to discuss the contents of the reports and assess the perceived usefulness of the accompanying recommendations. Their school will also be visited once per term (during implementation) by Darsel staff to provide such support in person. While T2 requires additional effort by teachers, it remains highly scalable. Darsel estimates T2 at national scale would require an additional 30 full-time equivalent (FTE) of staff. If T2 proves to be cost-effective, Darsel will scale it nationally. Comparisons between T1 and T2 will be particularly important for illustrating how much scaffolding is needed to make EdTech innovations effective or cost-effective.

2.2 Hypotheses

We formulate our hypotheses using the theory of change developed in Appendix section A. The academic year started in September 2025, and the Darsel intervention was implemented within one month of the start of the academic year, as soon as baseline data collection was completed. Our outcome measures will come from a midline assessment and student survey conducted midway through the academic year (late November/early December 2025), and an endline assessment and student survey conducted at the end of the academic year (May 2026). Teacher outcomes will come from midline and endline teacher surveys. Engagement (usage) will be measured using Darsel administrative data from the study period.

We will present all outcomes and test all hypotheses at midline and endline separately. We discuss below specific, additional hypotheses about the growth in learning and/or persistence in gains when comparing midline and endline. All hypotheses will be tested with the significance level (α) = 0.05.

To test the core assumption of the theory of change that usage of Darsel increases math achievement, we will conduct a mathematics assessment consistent with the Jordanian curriculum. The assessment will be designed in collaboration with the Ministry of Education to cover pre-requisite skills (grade 4/5 mathematics skills) and grade 6 mathematics skills from each unit. We will also draw on the 2011 TIMSS mathematics question bank for grade 4 to provide additional internationally validated questions. Table 2 in Appendix B shows the grade 6 Jordanian curriculum and Table 3 in Appendix B shows the planned assessment item types.

In line with psychometric measures of educational learning, we will transform the assessment responses into a standardized metric of learning using Item Response Theory (IRT), specifically a two-parameter logistic model (Jacob and Rothstein 2016). The underlying data for the IRT will be binary (correct = 1, incorrect or no answer = 0) for the mathematics assessment questions. We will thus treat missing responses as incorrect for the exam. For IRT, there will be some common (anchor) items in each of baseline, midline, and endline assessments, as shown in Table 3 in Appendix B, as well as an emphasis on pre-requisite skills for baseline, semester one skills for midline, and semester two skills for endline. Items will also vary in difficulty level. We will pilot all items to identify and remove any that perform poorly (e.g., floor or ceiling effects). To place items from each wave on a common scale, we will undertake IRT after endline data collection so that all items are on the same scale.

We will assess the impact of the treatment on the mathematics IRT score as our primary outcome. Therefore, our first hypothesis is:

H1: Darsel will increase students' mathematical skills, which will be measured by the overall mathematics latent value (θ) generated by our 2-parameter IRT model.

As a robustness check, we will also test whether final mathematics grades in the Ministry of Education's Education Management Information System (EMIS) are impacted by Darsel. Grades are per subject (math, science, English, Arabic) and based on first midterm, second midterm, participation score, and final exam. Exams are written by and grades given by teachers, although teachers may sometimes share the same exam within a school. Grades are on a 0 - 100 scoring system.

To test whether Darsel's motivational messages, gamification, and teaching at the right level lead to improvements in (math) confidence, motivation, and self-efficacy, we will measure students' socio-emotional skills. Specifically, we will use the brief form math self-efficacy scale and the brief form math anxiety scale, along with two items from the liking of math scale. These socio-emotional skill measures were designed and tested for United States 5th - 8th graders (Sinclair et al. 2025). An additional item on fixed versus growth mindset is included (Yeager et al. 2019). All these items will have a response scale of: (1) strongly disagree, (2) disagree, (3) agree, (4) strongly agree. Anxiety and fixed mindset items will be reverse coded for analyses. Furthermore, four utility-value items (oriented towards careers), designed and tested for grades 7 - 9 are included (Fiorella et al. 2021).[4] These have responses of (1) never, (2) rarely, (3) usually, (4) sometimes, (5) always. When students do not select a response, their outcome will be missing for the socio-emotional item in question. For outcomes that are indices, we will create a dummy for a missing item to include in the factor analysis, and set the missing item to the control mean for

[4] On feedback from MoE and the team in training workshops, we updated one of the four items from "I think about how learning math can help me get a good job" to "I think learning math can help me get a good job" on an agree/disagree rather than frequency scale.

the factor analysis. Observations with an entirely missing outcome will be dropped from the analysis of that outcome in that wave, but can be included for all other outcomes and waves (e.g., can be missing growth mindset at midline but still included for endline growth mindset and midline math anxiety, so long as those outcomes have data). If more than 5% of individuals are missing a specific outcome, we will undertake an analysis of this outcome akin to our attrition analysis to examine whether the missing data are missing at random.

We will present analyses for each category of outcomes: self-efficacy, anxiety, liking, fixed vs. growth mindset, and utility-value of math, as well as an overall socio-emotional index. The categories with multiple items and the overall socio-emotional index will be constructed via factor analysis. We will present, in the appendix, and briefly discuss in the paper, a validation of the factor analyses along a number of dimensions. We will include in the appendix the factor structures for all factors (Eigenvalues, uniqueness, factor loadings, scoring coefficients). We will also undertake, in the appendix, validation in terms of predictive validity at baseline; e.g., whether self-efficacy and anxiety are predictive of baseline math skills. For the control group, we will estimate the correlation of the factors from baseline to midline and endline, as a measure of reliability (albeit one that we would not in any case expect to be perfectly correlated over time). Furthermore, we will test for measurement invariance in the control group across baseline, midline, and endline in the appendix. Specifically, we will: (1) test for whether control group factor values are significantly different by wave for each factor; and (2) re-run the factor analysis separately by wave for the control group for each factor to see if the underlying factor structure varies appreciably, testing for significant differences in the scoring coefficients across waves by bootstrapping.

H2: Darsel will increase students' socio-emotional skills.

To test the downstream impact of these immediate outcomes, we will measure grade repetition and dropout rates based on the students' enrollment status and grade level in the following academic year (2026 - 2027) from the Ministry of Education's EMIS.[5]

H3: Darsel will decrease student dropout rates and grade repetition rates.

Dropout will be defined based on a student's national ID being enrolled per the EMIS in October 2026 (viz., the subsequent school year). We will treat a student being absent (missing) from the EMIS as an indicator of dropout. Grade repetition will be defined as a student's enrollment in

[5] We also hope to undertake a long-run follow-up of student test scores and grade progression in subsequent years using EMIS data but will do so in a subsequent paper in order to be able to contribute to policy discussions and decisions about AI and chatbots in Jordan, and globally, in a timely manner. By relying on EMIS data for future outcomes, future research would not require additional data collection and be low-cost.

grade 6 again in October 2026, per EMIS (i.e., grade 6 re-enrollment = 1, all other states including dropout = 0).

Our fourth hypothesis is that the treatment (especially T2) will increase teacher engagement, which will lead to further improvements beyond those of the main Darsel intervention (T1). To shed light on mechanisms, we will also interview grade 6 math teachers at baseline, midline, and endline.

H4: The additional support provided by T2 will lead to additional improvements in teaching and student outcomes, beyond those conferred by T1.

Specifically, we will test for differential treatment effects for T2 for math learning (H1), socio-emotional skills (H2), and grade repetition and drop out (H3). Our interpretation of the results of H4 will be informed by differences in student usage (see H6 below), as well as process monitoring data from teachers in T2 (described in the exploratory analyses).

Additionally, we test for changes in teacher effort and behaviors (H4). Specific outcomes for teachers are adapted from the TIMSS 2023 teacher questionnaire (International Association for the Evaluation of Educational Achievement (IEA) 2022) and will be:

Instructional quality: Sum of: How often do you do the following in teaching grade 6 mathematics?

(4) Every or almost-every lesson (3) About half of lessons (2) Some lessons (1) Never

- Relate lesson to students' daily lives.
- Ask students to explain their answers.
- Communicate lesson goals/objectives.
- Set challenging exercises beyond instruction.
- Encourage classroom discussions.
- Link new content to prior knowledge.
- Ask students to choose their own problem-solving

Teacher capacity: Sum of: How much do you agree or disagree?

(1) Agree a lot (2) Agree a little (3) Disagree a little (4) Disagree a lot

- I believe there are too many students in the classes.
- I believe that too many students lack prerequisite knowledge/skills
- I believe there is too much material to cover.
- I believe there are too many teaching hours.
- I need more time to prepare for class
- I need more time to assist individual students.
- I feel too much pressure from parents.
- I find difficulty keeping up with curriculum changes.
- I believe there are too many administrative tasks.

Teaching at the right level (TaRL): Number of yes answers to:
When you notice some of your students are falling behind, what have you done (in the last school year)? (select all that apply)

1 = Group the students in the class according to level
2 = Provide individualized and targeted instruction
3 = Provide individualized homework
4 = Review concepts from earlier grades
5 = Assign extra worksheets or homework assignments
6 = Reach out to the parents/guardians
97 = Other (Specify)

We will also create an overall teacher quality index summing all three of these dimensions. Missing data on teacher quality will be handled in the same fashion as for student socio-emotional outcomes.

We anticipate heterogeneous effects along a variety of dimensions of the student's experience following the introduction of Darsel. Approximately three-quarters of Darsel users for grade 7 were girls, suggesting the intervention may particularly benefit this group. Socioeconomic status plays an important role in children's school success in Jordan (Hailat 2019), and may mediate Darsel impacts. Refugee children in Jordan do successfully enroll in school, but often struggle to succeed and progress (Krafft et al. 2022). Class size may impact the ability of teachers to provide individualized instruction, and thus the benefits of Darsel's model. Likewise, teacher quality may affect the benefits of both T1 and especially T2. Teaching at the right level, for example, may be most effective for those with weaker baseline skills (Banerjee et al. 2016). Overall, sex, socioeconomic status, status as a refugee from Syria, the size of the classroom, baseline teacher quality, and baseline skills are all determinants that should mediate the effect of this new technology.

H5: The effect of Darsel on student performance will differ by sex, socioeconomic status, refugee status, classroom size, baseline teacher quality, and baseline skills.

We will estimate heterogeneity by each of these dimensions for all the preceding hypotheses. See the heterogeneity analysis section for specifications of these measures.

Usage of education technology is an important mechanism in increasing performance. *Exposure* to education technology does not necessarily lead to *utilization* of the technology (Wenglinsky 1998).

H6: The effect of Darsel on mathematics skills will differ by usage.

We will estimate heterogeneity in impacts for above vs. below median usage in the pooled treatments. We will descriptively explore impacts along the entire distribution of usage, in case we can identify any potential points of increasing or diminishing returns. Missing usage data will be treated as zero usage (i.e., never logging in). As discussed in the exploratory analyses section, we will also investigate a variety of patterns of usage and predictors of usage for both students and teachers.

The combination of the midline and endline provides an important opportunity to test the persistence of impacts. Impacts could potentially diminish over time, both due to novelty and thus usage declining or to the fade out of learning.

H7: The impacts of Darsel will vary between midline and endline.

To test the persistence of effects, we will use IRT on term 1 assessment items (see Table 3, Appendix B) and assess whether midline and endline impacts on term 1 items versus baseline are equal (H7).

H8: There will be no crowd-out effects of Darsel on other types of learning.

Spending time on Darsel could substitute for other activities, including crowding out other types of learning, such as literacy or science. To test the potential for crowd-out, we will examine both studying behaviors and grades as outcomes. For the grades outcomes, we will use the 0 - 100 science, English, and Arabic final grades from the EMIS. For the studying behaviors, we will use the following outcomes:

- The number of minutes spent studying or doing homework outside of school yesterday, separately for science and Arabic.
- Number of responses (other than no one) to:
  Who usually helps you with your schoolwork? Check all that apply.
  a) Mother
  b) Father
  c) Siblings
  d) Other relatives
  e) Friends or classmates
  f) No one

Missing data on these outcomes will be handled in the same fashion as the student socio-emotional and teacher quality outcomes.

2.3 Identification Strategy

We assess the impact of Darsel's AI chatbot through a RCT. This RCT will evaluate two versions of the program for grade 6. First, a "business as usual" (T1) intervention that reflects the current model of the Darsel platform that has been used for grade 7 in Jordan. Second, we introduce an intervention (T2) adding teacher encouragement and pedagogical advice (to teachers) features to the platform to evaluate their (combined) effectiveness in comparison to the current model.

We implement two stages of randomization. Randomization occurs first at the school level and then at the classroom level (in some schools). We randomized at the school level within each district and school sex. We have seven districts, and schools are either all-female or all-male. There are 21 schools of each sex selected in each district, from which seven are assigned to pure control, seven schools assigned to T1, and seven schools assigned to T2. We randomly generated ten possible randomizations. We then used multinomial logits that estimate treatment group assignment as a function of average class size for grade six, the number of grade six students in the school for each school (both from the 2025 - 2026 EMIS), and a dummy for whether in 2025-2026 there was at least one teacher teaching multiple classrooms (to ensure balance for classroom level randomization, described below). In addition, we included district-sex fixed effects to control for stratification. We then selected the randomization with the lowest overall chi-squared value (the most balanced of our ten randomizations).

In treatment schools (T1 or T2) with more than one class taught by the same math teacher, we conducted an additional classroom-level randomization. Specifically, we randomly chose one class (for one randomly selected math teacher if more than one is teaching two or more math classes) to be excluded from treatment and serve as a control. This allows us to conduct additional analysis relying on classroom-level randomization. We undertook a similar process of ten possible randomizations on the classroom level, estimating a multinomial logit with class size, and selecting the model with the lowest overall chi-squared value.

As a result, control schools contain only untreated classrooms, while treated schools include treated and untreated classrooms if there is at least one math teacher who teaches more than one math class. At the start of the 2025 - 2026 school year, MoE provided information on the number of classes per teacher in the sampled schools to facilitate implementation of this second level of randomization. In the sample, 255 of the 294 schools have multiple classrooms taught by the same teacher.

To estimate the effect of treatment, we will rely on two types of specifications, described further below. In the first, which uses a school-level randomization specification, the control group consists exclusively of classrooms in control schools, where neither Darsel in its current form nor the proposed enhanced model have been introduced. The treatment group consists of treated classrooms in the treatment schools. Therefore, this specification exploits the school-level randomization by comparing treated schools to control schools and excluding the (randomly) untreated classrooms in the treatment schools.

One concern with our approach could be spillovers from Grade 7, where treatment is universal, to Grade 6. Grade 6 teachers in these control schools may know of Darsel and may have been exposed to it through the rollout of Darsel aimed primarily at grade 7 students during the previous years. Despite the current presence of Darsel for higher grade students in this environment, grade 6 students are unlikely to have used or been encouraged to use Darsel through their school. Moreover, the publicly-available version of Darsel contains only the grade 7 content, which is more difficult than the grade 6 content and thus is unlikely to help grade 6 students. To assess the risk of spillovers, as well as non-compliance, we will collect data on familiarity and usage of Darsel at endline for teachers and students in both the treatment and control groups.

In the second specification, we exploit the randomization at the classroom level within those treatment schools with math teachers teaching multiple classes in grade 6. This approach allows us to include teacher fixed effects in the estimation to purge any teacher-specific factors from the treatment effect (in practice, we include school fixed effects, which are equivalent). This approach will allow us to isolate the impact of students using Darsel from any changes in teacher behavior that do not differ across treatment and control classes in the same school.

2.4 Statistical Power

Based on the 0.21 standard deviation effect size for the math assessment outcome estimated using pilot RCT data from Darsel (discussed below), the research team chose to target a minimum detectable effect (MDE) of 0.20, corresponding to common effect sizes found in the EdTech literature (Rodriguez-Segura 2022). We estimate this cluster design power calculation for the **school level randomization**, distinguishing (not pooling) T1 and T2, with the following assumptions. We used 80% power and 5% significance levels. We estimated the average number of grade 6 students per school to be 88 overall and 73 after excluding one class in the treatment schools with multiple classes, based on the data in our planned sample locations. Using the pilot RCT data we estimated an intraclass correlation of 0.362. We will gather baseline assessment data. There is a 0.644 test score correlation for students finishing basic education (grade 10) and secondary education (grade 12) from the 2016 Jordan Labor Market Panel Survey (JLMPS), which we used as a proxy for baseline and midline/endline assessment correlation, since pilot RCTs collected only endline data. Sampling clusters (schools) will be stratified at the district level and by schools' sex (boys only, girls only).

To calculate our desired MDE, we used Stata version 16.1, command clustersampsi. To obtain that MDE, we need 261 clusters (schools) and correspondingly 19,053 observations (students), split equally across the three arms (control and two treatments).[6] In this context, the research plan is for a sample of 294 schools, sampling the largest schools (in terms of numbers of students) to maximize power and reduce implementation logistics, with an estimated 21,462 grade 6 students (excluding control classroom students in treatment schools). For this planned sample, under the same assumptions as above, our power is 0.85 to detect a 0.20 SD effect.[7] Alternatively, our MDE at 0.80 power is 0.19 SDs.[8]

For the **classroom level randomization**, among the 263 schools with more than one class,[9] we expect one third (~87) to be in each treatment arm. In the 87 T1 and 87 T2 schools with more than one class, we use the same assumptions as above and a class size of 25 students, on average. There will be one control class and one or more treatment classes in each school. Note that students are randomly assigned to classes, so at baseline, the school ICC and class ICC are the same. We estimate power for the classroom level randomization very conservatively: assuming that there are just two classrooms (one treatment and one control) per school, and without accounting for any additional power generated from the school fixed effects incorporated into this model, as discussed below. The estimation uses just two classrooms per school because: (1) there is no existing estimator that accounts for both baseline correlation and varying cluster numbers, and (2) we do not know the number of math classes per teacher. The power with this (very conservative) version of estimates is 0.79[10] and the MDE is 0.20 SDs.[11]

3 Data

3.1 Sample

The study population is grade 6 students (mostly aged 11 - 12) who attend public schools in Jordan. The sample will include Jordanians and non-Jordanians, such as Syrian refugees, who are 7% of public school students (Assaad et al. 2023).

3.1.1 School Selection

The Jordanian Ministry of Education provided a list of all schools in Jordan with data from the 2024 - 2025 EMIS. Schools are organized into districts (N = 42 districts). Given that logistics

[6] clustersampsi, samplesize mu1(0) mu2(0.2) beta(0.8) alpha(0.05) sd1(1) sd2(1) m(73) rho(.3619568) base_correl(0.6439)

[7] clustersampsi, power alpha(0.05) sd1(1) sd2(1) mu1(0) mu2(0.2) k(98) m(73) rho(.3619568) base_correl(0.6439)

[8] clustersampsi, detectabledifference alpha(0.05) sd1(1) sd2(1) mu1(0) k(98) m(73) rho(.3619568) base_correl(0.6439)

[9] These power calculations used the 2024-2025 EMIS data, where we estimated 263 schools; in the 2025-2026 data this turned out to be the 255 schools mentioned above.

[10] clustersampsi, power alpha(0.05) sd1(1) sd2(1) mu1(0) mu2(0.2) k(87) m(25) rho(.3619568) base_correl(0.6439)

[11] clustersampsi, detectabledifference alpha(0.05) sd1(1) sd2(1) mu1(0) k(87) m(25) rho(.3619568) base_correl(0.6439)

take place on the district level, we plan to work in seven districts. We identified schools in all districts that included Grade 6 students in 2024 - 2025. Mixed-sex schools (almost all schools are single sex by grade 6) and schools in refugee camps (which require security clearance to visit) were removed. We then narrowed the sample of districts to those that had at least 21 girls' schools and 21 boys' schools with grade 6 to ensure an adequate sample size of schools. We subsequently selected districts to be from a variety of governorates (first administrative geography), to include a number of schools with Syrian refugees, and where the district had a large number of students. The decision to target the seven districts was driven, in part, by identifying locations where MoE supervisors were available to help facilitate implementation, for instance in accompanying proctors to the 294 schools and providing introductions and permissions. We also checked Darsel usage data for grade 7 to ensure that the sample did not have above-average (instead, slightly below average) usage in 2024 - 2025. The resulting sample of schools includes 25,996 grade 6 students according to the 2024 - 2025 data.

A stratified sample of schools from within these seven districts was selected. Schools were stratified at the district level and by schools' sex. In each district, we include the 21 largest all-boys schools and the 21 largest all-girls schools, in terms of grade 6 student enrollments during the preceding (2024 - 2025) school year, from school administrative data. Each stratum (district × sex) of 21 schools was split evenly into the 3 groups (control, T1, T2).

3.1.2 Participant Recruitment

We draw study participants from all grade 6 classes within the 294 selected schools. Jordanian grade 6 students are assigned one classroom that they stay in (with the same students) all day, while the teachers rotate by specialization through the classrooms over the course of the day. For our sample, we invite all grade 6 students within the selected schools to participate in the study. Only grade 6 mathematics teachers in the selected schools will be recruited to participate in the study and provide necessary support.

3.1.3 Consent and Ethical Considerations

We address ethical concerns through informed consent and steps to anonymize collected data. Informed consent will be obtained from teachers at the start of the teacher survey. For the student survey and assessment, the schools will reach out to parents or guardians to provide the opportunity for students to opt out. In addition, as part of the script at the start of in-school data collection, students will be informed about the study and that they can decline to answer the survey, assessment, or any specific question. Students and teachers can thus refuse to participate in any of the assessments or surveys. While identifiers are required to link the various data sources, we plan to securely store identifiable data in an encrypted format. All data will be de-identified before sharing with the PIs. Data made publicly available for future replication packages will be de-identified, and data access will require users to accept a license promising not to re-identify any schools, teachers, or students.

3.2 EMIS Data

Jordan's MoE has a cooperation agreement with Darsel that allows it to provide EMIS data to inform sampling and allow for the measurement of grade repetition and dropout outcomes. EMIS data from the 2024 - 2025 school year will be used at the end of the summer of 2025 to assess which schools are likely to have large numbers of students in grade 6 in the 2025 - 2026 school year. EMIS data will be used to identify the largest schools within the district for sampling. EMIS data from the fall of 2026 will be used to create outcomes for grade repetition and student drop out.

3.3 Darsel Administrative Data

During the intervention, Darsel will track data on students' and teachers' engagement with the platform (by tracking their usage). Students will use identifiable codes, which can be linked back to their national ID, to access the platform. Each interaction students have with the platform is recorded, including the start date, start time, duration, questions attempted (which are mapped to different topics, difficulty level, etc.), and answers' accuracy. We will thus be able to generate variables measuring the different iterations of usage based on the frequency and duration of interactions with the student chatbot.

Teachers' engagement with reports will be tracked. For both T1 and T2, teachers receive a standard report by sending a WhatsApp message of "report" to Darsel's chatbot. Each report request is thus logged and will be our main measure of teacher engagement. The team will also track the number of messages sent by teachers on the district-wide WhatsApp groups, including inbound support requests (through direct WhatsApp messages or phone calls to the support team), and responses to proactive outreach to teachers (for T2 only). For calls, duration will be logged. Attendance at virtual (T1 and T2) and in-person (T2) training sessions will also be recorded. We will report means and medians of all these measures descriptively, by treatment arm, when discussing take-up.

3.4 Survey Data

Measurement instruments consist of student surveys, student assessments, and teacher surveys. All instruments will include the respondent's national ID variable to ensure data linking. These instruments were developed jointly with experts from Jordan's MoE and subject-matter experts. For example, we consulted a global socio-emotional learning expert in designing those items. Jordan's MoE provided different question options for the math assessment, which were reviewed and then piloted, with IRT of the pilot data being used to select the better performing questions.

Instruments that were initially in English were translated into Jordanian Arabic by Arabic speakers, and we had a multitude of native Jordanian Arabic speakers review the instruments. All instruments were further reviewed before piloting as part of a joint J-PAL and Jordanian MoE week-long workshop, which included 20 MoE staff. The Jordanian MoE staff provided extremely detailed comments on question design, for instance on the number of categories in a Likert scale Jordanian students are used to, or clarifying wording such as "I believe" about growth mindset statements. Our Jordanian data collection firm also reviewed and provided feedback from their perspective and a student pilot of the questionnaires. We are therefore quite confident in the translation and contextual appropriateness of all questions. For all the surveys, proctors will be blind to treatment assignment. Both treatment and control students will be asked about Darsel use at endline in order to ensure blinding and to measure potential control group contamination.

Students' data will be collected through paper assessments and questionnaires filled in by the students under the supervision of proctors who will be hired from an independent data collection firm. At baseline, students were given a short (15 minutes) survey questionnaire to measure their demographics and socio-emotional skills. At midline and endline the (time invariant) demographic questions will be dropped but otherwise the same survey will be repeated. At baseline, endline, and midline, after completing the short survey, students will be given 60 minutes to complete the mathematics assessment. For consistency, timing will be strictly controlled by the proctors.

Both the short survey and the student assessments will have their results coded by an AI-based autograding system. Jordanian students are used to exams where they circle responses, so paper assessments and surveys with multiple choice responses will be used and students will circle their responses. These paper instruments will have their national ID. These instruments will then be scanned into PDF format and subsequently coded into a CSV file of responses. The resulting data will be student level observations for each wave, with variables for each survey question denoting the response (e.g., student 1 answered multiple choice response B for question 3). Each response is independently assessed by a commercially-available LLM based on the corresponding section of the scanned page. Darsel has piloted using this system and found that it was >99% accurate when manually checking 5% of responses. Error cases were all unclear situations, for instance, when the student selected multiple responses to a question, erasing one partially, and the AI made a questionable decision about the likely true answer, which would be an issue with human coders as well. The J-PAL MENA team will be provided a copy of the AI tool and undertake the data processing to preclude any potential conflicts of interest if Darsel did the data processing; the AI tool will be blind to students' treatment status.

The J-PAL MENA team will undertake quality control of the AI tool, checking 3% of the tool's coding by hand and reporting item-level and overall statistics (which will be provided in an appendix) on its coding accuracy at baseline, midline, and endline, overall and (after blind coding) by treatment status. Furthermore, we will undertake analyses that compare human-human inter-rater reliability to AI-human inter-rater reliability by having 3% of the manually entered responses coded by two different research assistants. Item-level and overall human-human inter-rater reliability will be presented in the same appendix table as AI-human inter-rater reliability to compare their relative performance.

For the teachers' data, an online survey on demographics and teaching practices will be designed and programmed by the researchers in SurveyCTO to be filled by the teachers through a link that is sent to them via WhatsApp. The proctors will share the link and encourage teachers to complete the survey while visiting the school and will follow up with any teachers who have not or who were absent that day to ensure completion.

Baseline data collection took place in the beginning of the academic year in Jordan (September 2025). A midline survey will be conducted in late November/early December 2025 (end of first semester). Similarly, the endline survey will be conducted at the end of the academic year (May 2026).

3.4.1 Non-response and attrition

All individuals who are in sample schools, even teachers or students who did not comply with the treatment or complete the baseline survey, will be eligible for subsequent midline and endline surveys in the RCT. If non-response of students at baseline is more than 5% we will use EMIS administrative data on the characteristics of students who did not respond to the baseline survey to check for non-random non-response by treatment status, district, and student demographics (sex, nationality, and age in months at the start of the school year).

There is potential for attrition from the baseline survey to the midline and endline surveys. We will allow students and teachers who did not respond at the midline or baseline to still participate in the endline date collection. We will document rates of attrition, as well as test for non-random attrition by treatment arm at midline and endline (separately). If attrition reaches more than 5% for either midline or endline we will also model attrition with covariates of baseline outcomes (primary and secondary) and the dimensions of heterogeneity detailed in section 4.4. In addition, as a sensitivity analysis for our results, we will estimate Lee bounds, testing how our results would change assuming, variously, the best and worst outcomes for attritors and similarly for baseline non-respondents. We will present three checks: (1) standard Lee bounds (Lee 2009); (2) generalized Lee bounds (Semenova 2025), using the baseline outcomes (discussed above) and dimensions of heterogeneity (discussed below) as covariates; and (3) standard Lee bounds drawing from the best and worst performers in the same classroom, rather than the overall distribution.

3.4.2 Non-compliance

Non-compliance with treatment could occur if individuals assigned to the treatment do not use Darsel, or if individuals assigned to the control arm manage to access Darsel. We will have usage measures, as discussed above, from administrative data to detect never-takers, and access to the grade 6 platform will require an access code to help ensure compliance, which should make always-takers very rare. At endline, we will ask teachers and students whether they used Darsel at all, and specifically whether they used the Darsel content for grade 6; we will do so only at endline so as not to raise awareness of the platform in the control group and thus create the contamination problem we are trying to minimize. We will report descriptively the means for non-compliance by treatment arm for students and teachers, noting that, for teachers, use of Darsel for other grades may be part of normal practice and thus not non-compliance (e.g., if the math teacher teaches both grade 6 and grade 7).

3.5 Pilot Data

Darsel has been used by over 260,000 students since its launch in 2021, largely as a result of government partnerships in Jordan, India, and Nigeria. In Jordan, which was Darsel's first country of operations, Darsel was nationally adopted by the Ministry of Education for grade 7 in 2023, and it remains in use by all government schools.

Darsel is on similar growth trajectories in the two other countries. In Nigeria, Darsel has been working with Lagos State since 2023, and its program was adopted for grade 9 in all government schools in the 2024-2025 academic year, reaching an estimated 30,000 students. The Lagos program is being expanded to include grades 7 and 8 in 2025-2026, and planning is also underway for pilots in 3 other Nigerian states. In India, Darsel previously conducted pilots in both Delhi and Rajasthan, and recently signed memoranda of understanding (MoUs) in Haryana and Karnataka, paving the way for state-wide adoption.

Darsel has undertaken a number of internal pilot RCTs of its chatbot, including a 2022 pilot RCT for grade 7 and two pilot RCTs in 2025 for grade 4 and grade 6, all in Jordan. The 2022 RCT, which took place in 10 schools and randomized on the classroom level across 29 grade 7 classrooms, found a 0.16 SD effect size ($p = 0.030$). Pilots in other countries were not randomized.

We report here - and used for our power calculations - the 2025 pilot RCT for grade 6 students. This pilot RCT included 40 schools, 93 classrooms, and 2,416 students, with randomization at the classroom level. The schools were selected based on nomination by the district supervisors, who sought to maximize uptake by selecting schools with collaborative principals and teachers, and/or a history of high usage in Darsel for grade 7. As such, the selection of schools is not representative, and is explicitly biased towards higher-performing schools. Students were given a curriculum-aligned math assessment developed by MoE at endline. There was no baseline. The outcome was the normalized total number of questions correct (total correct ranged from 0 to 28; mean = 11.7; SD = 5.2). The treatment was standard Darsel (T1). Intention to Treat (ITT) estimates were undertaken, clustering standard errors at the classroom (randomization) level. No controls were included in the model specification. We report the results of two specifications: (1) simply regressing treatment on the normalized total number of questions correct, and (2) including school fixed effects (estimated using reghdfe, version 6.12.3 in Stata 16.1). The impact of Darsel is 0.199 SDs in specification (1), p = 0.174, 95% CI of -0.089 to 0.487. In specification (2) with the school fixed effects, the impact is 0.213 SDs, p = 0.001, 95% CI of 0.083 to 0.342. We use the school fixed effects estimate for our power calculations, above. The pilot results illustrate both the potential impact of Darsel and the need for a better powered study in order to examine, for example, heterogeneous effects.

4 Analysis

4.1 Statistical Methods

Our identification strategy for estimating the impact of the Darsel program is based on random assignment of schools to treatment arms T1, T2, or control. This will allow us to use ordinary least squares (OLS) to estimate the causal effects of being assigned to a T1 or T2 school. We will estimate both the ITT effect of the program and the local average treatment effect (LATE), based on usage. Since it is very unlikely that students in control schools will gain access to Darsel's grade 6 materials, LATE is the average treatment effect on the treated (ATT).

4.2 Statistical Model

The evaluation will test whether a student's use of Darsel directly leads to increased math and socio-emotional skills. Accordingly, we test our primary hypotheses of improved outcomes via a pooled regression (combining both treatments). For our primary ITT estimates (H1 - H3, H5 - H8), we will exploit randomization at the school level, comparing treated schools to control schools, but dropping the control classrooms in treated schools from our sample:

$$(1)\ Y_{isr} = \alpha_r + \beta T_{sr} + \boldsymbol{\gamma}'\mathbf{X}_{isr,t=0} + \varepsilon_{isr}$$

where Y_{isr} is the outcome of interest for student i in school s in strata r, α_r is a fixed effect for strata r (strata are seven districts and school type (boys or girls) within each district), T_{sr} is an indicator the school s in strata r is a treatment (T1 or T2) school, $\mathbf{X}_{isr,t=0}$ is a vector of control variables that are used to increase precision (discussed in the following section), where the $t = 0$ subscript indicates that they are measured at baseline, and ε_{isr} is a residual term that is uncorrelated with T_{sr} since schools are randomly assigned to treatment. Standard errors will be clustered at the school level for the specifications relying on school-level randomization since this is the level of treatment assignment (Abadie et al. 2023). The coefficient of interest is β, the estimate of the ITT effect of the Darsel program.

To separate out the effects of the different treatments (T1 vs. T2, estimated for H1, H2, H3, H4, H5, H7) we will estimate:

(2) $Y_{isr} = \alpha_r + \beta_1 T1_{sr} + \beta_2 T2_{sr} + \boldsymbol{\gamma}'\boldsymbol{X}_{isr,t=0} + \varepsilon_{isr}$

The coefficients of interest, β_1 and β_2, will be tested for equivalence.

The alternative empirical specification exploits randomization at the classroom level (testing the same hypotheses as in the main specifications, except for heterogeneity [H5]). It restricts the sample to the treated schools and classrooms taught by the teacher that had one classroom randomized to be untreated and the rest treated, with the control group consisting of control classrooms in treated schools.

(3) $Y_{isc} = \alpha_s + \beta T_{sc} + \boldsymbol{\gamma}'\boldsymbol{X}_{isc,t=0} + \varepsilon_{isc}$

Where Y_{isc} is the outcome of interest for student i in school s and classroom c, a_s is a fixed effect for school s, T_{sc} is a treatment indicator for the classroom c in the school s, $X_{isc,t=0}$ is a vector of control variables that are used to increase precision, and ε_{isc} is the error term, which is uncorrelated with T_{sc} since classrooms will be randomly assigned to the Darsel program. Standard errors will be clustered at the classroom level following Abadie et al. (2023). The coefficient of interest is β, the estimate of the ITT effect of the Darsel program within schools, comparing treatment and control classrooms.

Similarly, we will estimate the model separating the effects of T1 vs. T2 using the classroom level randomization:

(4) $Y_{isc} = \alpha_s + \beta_1 T1_{sc} + \beta_2 T2_{sc} + \boldsymbol{\gamma}'\boldsymbol{X}_{isc,t=0} + \varepsilon_{isc}$

Again, the coefficients of interest, β_1 and β_2, will be tested for equivalence.

For all LATE estimates, we define a variable of ever-usage ("usage") of Darsel. The two estimating equations for LATE are the following:

(5) $Usage_{isr} = \lambda_r + \tau T_{sr} + \boldsymbol{\pi}'\boldsymbol{X}_{isr,t=0} + \omega_{isr}$

(6) $Y_{isr} = \delta_r + \theta \widehat{Usage}_{isr} + \boldsymbol{\rho}'\boldsymbol{X}_{isr,t=0} + \eta_{isr}$

where all variables are as defined above, λ_r and δ_r are strata fixed effects, and η_{isr} and ω_{isr} are residual terms that are uncorrelated with T_{sr} since schools are randomly assigned to the Darsel program. The coefficient of interest is θ, the estimate of the LATE of the Darsel program. Residuals will be clustered at the school level in both equations. Because we are examining usage, which may be driven by T1 vs. T2, we will estimate only pooled models for LATE.

4.3 Controls

The baseline controls expressed by $X_{isr,t=0}$ in section 4.2 will be (1) math assessment scores, and (2) values of the socio-emotional index. We will average these at the teacher level for teacher outcomes. Controls will also include any characteristics that are imbalanced at baseline, as discussed below in the section on baseline balance.

When modeling the sensitivity analysis for student final grades from the EMIS, we will also include average final grades from their teacher the preceding year, or a control for whether it is the teacher's first year (and thus average final grades would otherwise be missing). If any controls are missing, we will create an indicator for that control being missing, include that indicator in our models, and replace the missing value with 0s (if the control is binary or categorical) or the mean of the control group (if the control is a count or continuous variable).

4.4 Multiple Outcomes and Multiple Hypothesis Testing

As we test many hypotheses, the risk of falsely rejecting at least one null hypothesis increases simply because of the number of tests being conducted. For example, if we test 20 independent hypotheses at the 5% level, the chance of wrongly rejecting at least one is 64%. To guard against this, we adjust for multiple hypothesis testing. The simplest method is to apply the Bonferroni bound, which simply increases the significance level from α to α/m, where m is the number of hypotheses tested. This approach is, however, overly conservative and results in low power. We therefore follow Benjamini et al. (2006), who improve on Bonferroni by using resampling to approximate what the distribution of p-values would look like if all null hypotheses were true, and then compare our observed p-values to that benchmark. We follow the implementation in Anderson (2008) and will report for each hypothesis the adjusted p-value ("sharpened q-value"), providing a transparent measure of significance after adjustment.

We define eight families of hypotheses, each of which is adjusted separately. The first four families contain 12 hypotheses each (covering pooled treatment, treatment 1, treatment 2, and treatment 1 vs. treatment 2, each across our three main outcomes (indices for math assessment, socio-emotional skills, and teacher quality), time points (midline and endline), and samples (school vs. classroom randomization levels)). The second four families contain 6 hypotheses each (covering whether effects change over time within each treatment comparison). A summary of the eight families is shown in Table 1 below.

Table 1. Families of hypotheses

Family	Hypotheses included	No. of hypotheses
1	ITT effect of pooled treatment vs. control on 3 outcomes × 2 time points × 2 samples	12
2	ITT effect of treatment 1 vs. control on 3 outcomes × 2 time points × 2 samples	12
3	ITT effect of treatment 2 vs. control on 3 outcomes × 2 time points × 2 samples	12
4	ITT effect of treatment 1 vs. treatment 2 on 3 outcomes × 2 time points × 2 samples	12
5	Change over time (midline vs. endline) for pooled treatment vs. control on 3 outcomes × 2 samples	6
6	Change over time (midline vs. endline) for treatment 1 vs. control on 3 outcomes × 2 samples	6
7	Change over time (midline vs. endline) for treatment 2 vs. control on 3 outcomes × 2 samples	6
8	Change over time (midline vs. endline) for treatment 1 vs. treatment 2 on 3 outcomes × 2 samples	6

4.5 Heterogeneous Effects

Heterogeneity analyses will focus only on the three indices (math assessment, socio-emotional skills, and teacher quality) and assess impacts for student outcomes by: (1) sex, (2) refugee status, (3) tertiles of socioeconomic status, (4) tertiles of class size, (5) tertiles of baseline teacher quality, (6) tertiles of baseline math scores, and (7) tertiles of socio-emotional index. For teacher outcomes we will assess heterogeneity by: (1) tertiles of their average class size, (2) tertiles of baseline teacher quality, and (3) tertiles of their students' baseline math skills. We will specifically estimate equations 1 (pooled treatment) and 2 (T1 vs. T2) for subgroups and then test for equivalence of the coefficients of interest.

The data for heterogeneity analysis subgroups will be as follows:

1) We will estimate heterogeneity analyses by sex (male vs. female). EMIS provides data on students' sex.

2) We will undertake heterogeneity analyses by refugee status, comparing Jordanians to Syrians – excluding other nationalities for this analysis. Refugees will be defined as individuals whose nationality is Syrian in the EMIS administrative data. Although Jordan historically had some economic migrants from Syria, the vast majority (93%) of Syrians in the country are refugees (Krafft et al. 2019). Due to the logistical and permitting challenges of data collection in the refugee camps, schools in refugee camps will not be included in the sample. However, 87% of Syrians reside in host communities and only 13% in camps (Krafft et al. 2019). Syrians have been integrated into the Jordanian public education system, although in some cases they attend second shift schools (Krafft et al. 2022), which are identified as separate schools by the EMIS and in our data.
3) Socioeconomic status tertiles will be based on tertiles of an index from a factor combining asset questions, and mother's and father's education level questions from the baseline student survey. Assets are from a question series of "Does your family own any…" with yes/no responses for computer, mobile phone, television, microwave oven, car, vacuum, freezer, and dishwasher. These assets were selected based on the JLMPS 2025 preliminary data as having a reasonable degree of variation.
4) Tertiles of class size will be calculated using the sampled schools based on EMIS enrollment data.
5) Tertiles of baseline teacher quality will be from the sum of the baseline instructional quality, teacher capacity, and teaching at the right level measures.
6) Estimates will be undertaken separately for tertiles of (a) baseline mathematics assessment and (b) baseline socio-emotional skill index.

4.6 Exploratory Analyses

Additional analyses will be exploratory in nature and presented primarily in appendices. For example, we will:

- Examine impacts by math skill level or curriculum unit, undertaking IRT separately for each unit or skill level's questions. Impacts will be estimated overall (LATE) and with topic or skill-level-specific usage for the LATE (example: impact of geometry practice on Darsel on geometry scores). We will specifically estimate separate impacts on pre-requisite skills (grade four/five and TIMSS items), semester one units, and semester two units, as well as individual units (e.g., fraction operations, see Appendix B, Table 2). We will also estimate separately for each lesson number (e.g., lesson 1, lesson 2) across units.
- Examine potential complementarity or substitutability of math skills when exploring impacts by curriculum unit, adding usage by *different* topics (example: impact of geometry practice in Darsel on algebra scores).
- Investigate impacts of Darsel on students' math skills at the item level, specifying the exact lesson and proficiencies covered by the item.
- Investigate impacts of Darsel on teaching at the right level at the item level for each of the TaRL items.

- Assess the quality of chatbot responses, following a process inspired by Björkegren et al. (2025). Specifically, we will randomly sample 250 student interactions with the Darsel chatbot for grade six and validate them in two stages. First, we will provide 25 Jordanian grade six math teachers who are not in the study sample with 10 different initial interactions (e.g., a request for a practice problem on fractions). We will then ask these teachers to provide what their response would be for each of the 10 interactions. Second, an additional 25 Jordanian grade six math teachers who are not in the study sample and did not provide initial responses will be asked to score a random sample of 10 of the chatbot responses from the actual interactions and a random sample of 10 teacher responses to different questions from the same 250 interactions. They will score the responses of the chatbot and the responses of the first group of teachers in terms of their relevance and helpfulness. The total of 50 teachers will be randomized into whether they provide the initial response or rate the responses. The second set of teachers will be blind to whether they are evaluating a human or chatbot response. We will then assess whether Darsel chatbot responses are rated significantly differentially than teacher responses on the two outcomes of relevance or helpfulness.
- Collect and analyze process monitoring data from teachers in T2 to explore potential behavioral mechanisms. We will specifically ask, via a Darsel WhatsApp poll:
  - Does the Darsel report help you know or track your students' performance levels in different topics? (Yes/no)
  - Have the performance data in the Darsel reports, such as collective and individual mastery levels for each lesson, or review suggestions, affected the teaching process in any way? For example, have you ever decided to review a particular topic because of the report? (Yes/no)
  - Can you give an example of how the report affected your teaching behavior? And what motivated that? (Open ended)
  - What feedback or comments do you have on the Darsel reports? How can we make them more useful? (Open ended)
  - We undertook this poll already during fall semester. There was a 3% non-response rate. Among all T2 teachers (including non-response), 89% of all T2 teachers answered "Yes" to Q1 (reported that Darsel helped them track performance), and 86% answered "Yes" to Q2 (reported that their behavior had been affected).
  - At endline, we will also ask about Darsel usage from all teachers
- Examine heterogeneous impacts by teacher characteristics, such as years of experience and degree/specialization, that are of particular interest to MoE.
- We will further explore how impacts vary from midline to endline.
- We will descriptively explore impacts along the entire distribution of usage, in case we can identify any potential points of, for example, diminishing returns.

- Investigate lesson-level usage (student selection of different topics) on the extensive and intensive margins to compare the impact of exposure to content to the impact of increasing usage (dosage) on a topic.
- Regular engagement is critical in EdTech interventions, and usage of EdTech often varies substantially across students (Muralidharan et al. 2019), suggesting that usage is a key mechanism for estimating the impact. Therefore, we will treat engagement as a secondary outcome, explored both descriptively and in terms of predictors. Specifically, we will examine engagement of students and teachers on both the extensive margin (binary for any usage) and intensive margin (active weeks, active days for both teachers and students, and for students duration of total use, number of questions attempted). We will also create a duration outcome, time of first usage duration, in days, to last usage, for both teachers and students' engagement with the platform. We will analyze these outcomes in a variety of ways:
    - Descriptively, we will explore the mean of these engagement outcomes by treatment arm and each of our dimensions of heterogeneity.
    - We will plot retention curves using a Kaplan-Meier survival function to estimate the probability that students or teachers remain active on the Darsel platform over time, beginning with their first use.
    - Furthermore, we will model usage (as defined above) as a secondary outcome, using a model similar to our LATE estimate first stage (equation 5). This model will be restricted to treatment classes in treatment schools. We will estimate the effects of T1 vs. T2 by including a control for T2 and interacting T2 with all covariates, as in:

      $$(7)\ Usage_{isc} = \alpha_s + \beta_1 T2_{sc} + \boldsymbol{\gamma}'\mathbf{X}_{isc,t=0} + \boldsymbol{\pi}'T2_{sc}{*}\mathbf{X}_{isc,t=0} + \varepsilon_{isc}$$

      This will allow us to understand who is most likely to use Darsel and who has particularly shifted into Darsel. Specific covariates will include:
        - All of the dimensions of heterogeneity
        - Teacher engagement (for student engagement only),
        - Whether the family owns a mobile phone
        - Baseline minutes spent studying mathematics yesterday
        - Minutes spent using a phone, computer, of the internet for activities not related to school at baseline
        - Dummies for whether any of the following helps with school work at baseline: mother, father, siblings, other relatives, friends or classmates
        - Teacher gender
        - Teacher number of years teaching at baseline
        - Teacher highest level of education completed at baseline

    - Teacher participation in professional development within the past two years (at baseline)
- Compare Darsel’s cost-effectiveness per 0.1 SD learning gain to the EdTech literature and evidence on the impact of other learning interventions in LMICs (e.g., Muralidharan et al. 2019; Kremer et al. 2013; Angrist et al. 2025). See Appendix C for details.

5 Bibliography

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. 2023. “When Should You Adjust Standard Errors for Clustering?” *The Quarterly Journal of Economics* 138 (1): 1–35. https://doi.org/10.1093/qje/qjac038.

Anderson, Michael L. 2008. “Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects.” *Journal of the American Statistical Association* 103 (484): 1481–95. https://doi.org/10.1198/016214508000000841.

Angrist, Noam, Peter Bergman, and Moitshepi Matsheng. 2022. “Experimental Evidence on Learning Using Low-Tech When School Is Out.” *Nature Human Behaviour* 6 (7): 941–50. https://doi.org/10.1038/s41562-022-01381-z.

Angrist, Noam, David K. Evans, Deon Filmer, Rachel Glennerster, Halsey Rogers, and Shwetlena Sabarwal. 2025. “How to Improve Education Outcomes Most Efficiently? A Review of the Evidence Using a Unified Metric.” *Journal of Development Economics* 172: 103382. https://doi.org/10.1016/j.jdeveco.2024.103382.

Assaad, Ragui, Thomas Ginn, and Mohamed Saleh. 2023. “Refugees and the Education of Host Populations: Evidence from the Syrian Inflow to Jordan.” *Journal of Development Economics* 164: 103131. https://doi.org/10.1016/j.jdeveco.2023.103131.

Bandura, Albert. 1977. “Self-Efficacy: Toward a Unifying Theory of Behavioral Change.” *Psychological Review* 84 (2): 191–215. https://doi.org/10.1037/0033-295X.84.2.191.

Banerjee, Abhijit, Rukmini Banerji, James Berry, et al. 2016. “Mainstreaming an Effective Intervention: Evidence From Randomized Evaluations of ‘Teaching At the Right Level’ in India.” *NBER Working Paper Series* No. 22746.

Beg, Sabrin, Waqas Halim, Adrienne M. Lucas, and Umar Saif. 2022. “Engaging Teachers with Technology Increased Achievement, Bypassing Teachers Did Not.” *American Economic Journal: Economic Policy* 14 (2): 61–90. https://doi.org/10.1257/pol.20200713.

Benjamini, Yoav, Abba M. Krieger, and Daniel Yekutieli. 2006. “Adaptive Linear Step-up Procedures That Control the False Discovery Rate.” *Biometrika* 93 (3): 491–507. https://doi.org/10.1093/biomet/93.3.491.

Bettinger, Eric, Robert Fairlie, Anastasia Kapuza, Elena Kardanova, Prashant Loyalka, and Andrey Zakharov. 2023. “Diminishing Marginal Returns to Computer-Assisted Learning.” *Journal of Policy Analysis and Management* 42 (2): 552–70. https://doi.org/10.1002/pam.22442.

Björkegren, Daniel, Jun Ho Choi, Divya Panchaksharappa Budihal, Dominic Sobhani, Oliver Garrod, and Paul Atherton. 2025. “Could AI Leapfrog the Web? Evidence from Teachers in Sierra Leone.” arXiv:2502.12397. Preprint, arXiv, December 2. https://doi.org/10.48550/arXiv.2502.12397.

Bold, Tessa, Deon Filmer, Gayle Martin, et al. 2017. “Enrollment without Learning: Teacher Effort, Knowledge, and Skill in Primary Schools in Africa.” *Journal of Economic Perspectives* 31 (4): 185–204. https://doi.org/10.1257/jep.31.4.185.

Büchel, Konstantin, Martina Jakob, Christoph Kühnhanss, Daniel Steffen, and Aymo Brunetti. 2022. “The Relative Effectiveness of Teachers and Learning Software: Evidence from a Field Experiment in El Salvador.” *Journal of Labor Economics* 40 (3): 737–77. https://doi.org/10.1086/717727.

Cardim, Joana, Teresa Molina-Millán, and Pedro C. Vicente. 2023. “Can Technology Improve the Classroom Experience in Primary Education? An African Experiment on a Worldwide Program.” *Journal of Development Economics* 164: 103145. https://doi.org/10.1016/j.jdeveco.2023.103145.

Csikszentmihalyi, Mihaly. 1990. *Flow: The Psychology of Optimal Experience*. Harper & Row.

Cutler, David, and Adriana Lleras-Muney. 2012. “Education and Health: Insights from International Comparisons.” *NBER Working Paper Series* No. 17738. https://doi.org/10.3386/w17738.

Davier, M. von, A. Kennedy, K. Reynolds, et al. 2023. *TIMSS 2023 International Results in Mathematics and Science*. Boston College, TIMSS & PIRLS International Study Center.

De Simone, Martın, Federico Tiberti, Maria Barron Rodriguez, Federico Manolio, Wuraola Mosuro, and Eliot Jolomi Dikoru. 2025. “From Chalkboards to Chatbots: Evaluating the Impact of Generative AI on Learning Outcomes in Nigeria.” *World Bank Policy Research Working Paper Series* No. 11125.

Department of Statistics (Jordan). 2023. *Jordan in Figures 2022*. Amman, Jordan.

El-Kogali, Safaa El Tayeb, and Caroline Krafft, eds. 2020. *Expectations and Aspirations: A New Framework for Education in the Middle East and North Africa*. World Bank.

Escueta, Maya, Andre Joshua Nickow, Philip Oreopoulos, and Vincent Quan. 2020. “Upgrading Education with Technology: Insights from Experimental Research.” *Journal of Economic Literature* 58 (4): 897–996. https://doi.org/10.1257/jel.20191507.

Evans, David K., and Amina Mendez Acosta. 2021. “Education in Africa: What Are We Learning?” *Journal of African Economies* 30 (1): 13–54. https://doi.org/10.1093/jae/ejaa009.

Fiorella, Logan, So Yoon Yoon, Kinnari Atit, et al. 2021. “Validation of the Mathematics Motivation Questionnaire (MMQ) for Secondary School Students.” *International Journal of STEM Education* 8 (1). https://doi.org/10.1186/s40594-021-00307-x.

Gethin, Amory. 2025. “Distributional Growth Accounting: Education and the Reduction of Global Poverty, 1980–2019.” *The Quarterly Journal of Economics*, qjaf033. https://doi.org/10.1093/qje/qjaf033.

Hailat, Mahmoud Ali. 2019. "Education of Jordanians: Outcomes in a Challenging Environment." In *The Jordanian Labor Market: Between Fragility and Resilience*, edited by Caroline Krafft and Ragui Assaad. Oxford University Press.

Hattie, John, and Helen Timperley. 2007. "The Power of Feedback." *Review of Educational Research* 77 (1): 81–112. https://doi.org/10.3102/003465430298487.

Henkel, Owen, Hannah Horne-Robinson, Nessie Kozhakhmetova, and Amanda Lee. 2024. "Effective and Scalable Math Support: Evidence on the Impact of an AI- Tutor on Math Achievement in Ghana." Paper presented at International Conference on Artificial Intelligence in Education. *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*.

Howard-Jones, Paul A. 2010. *Introducing Neuroeducational Research: Neuroscience, Education and the Brain from Contexts to Practice.* Routledge.

International Association for the Evaluation of Educational Achievement (IEA). 2022. "IEA's Trends in International Mathematics and Science Study – TIMSS 2023." TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College.

Jacob, Brian, and Jesse Rothstein. 2016. "The Measurement of Student Ability in Modern Assessment Systems." *Journal of Economic Perspectives* 30 (3): 85–108. https://doi.org/10.1257/jep.30.3.85.

Kerwin, Jason T., and Rebecca Thornton. 2021. "Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures." *The Review of Economics and Statistics* 103 (2): 251–64. https://doi.org/10.1162/rest_a_00911.

Koval-Saifi, Nedjma, and Jan Plass. 2018a. *Antura and the Letters: Impact and Technical Evaluation*. World Vision and Foundation for Information Technology Education and Development.

Koval-Saifi, Nedjma, and Jan Plass. 2018b. *Feed the Monster: Impact and Technical Evaluation*. World Vision and Foundation for Information Technology Education and Development.

Krafft, Caroline, Maia Sieverding, Nasma Berri, Caitlyn Keo, and Mariam Sharpless. 2022. "Education Interrupted: Enrollment, Attainment, and Dropout of Syrian Refugees in Jordan." *The Journal of Development Studies* 58 (9): 1874–92. https://doi.org/10.1080/00220388.2022.2075734.

Krafft, Caroline, Maia Sieverding, Colette Salemi, and Caitlyn Keo. 2019. "Syrian Refugees in Jordan: Demographics, Livelihoods, Education, and Health." In *The Jordanian Labor Market Between Fragility and Resilience*, edited by Caroline Krafft and Ragui Assaad. Oxford University Press.

Kremer, Michael, Conner Brannen, and Rachel Glennerster. 2013. "The Challenge of Education and Learning in the Developing World." *Science* 340 (6130): 297–300. https://doi.org/10.1126/science.1235350.

Lai, Fang, Renfu Luo, Linxiu Zhang, Xinzhe Huang, and Scott Rozelle. 2015. "Does Computer-Assisted Learning Improve Learning Outcomes? Evidence from a Randomized Experiment in Migrant Schools in Beijing." *Economics of Education Review* 47: 34–48. https://doi.org/10.1016/j.econedurev.2015.03.005.

Lai, Fang, Linxiu Zhang, Xiao Hu, et al. 2013. "Computer Assisted Learning as Extracurricular Tutor? Evidence from a Randomised Experiment in Rural Boarding Schools in Shaanxi." *Journal of Development Effectiveness* 5 (2): 208–31. https://doi.org/10.1080/19439342.2013.780089.

Lally, Phillippa, Cornelia H. M. Van Jaarsveld, Henry W. W. Potts, and Jane Wardle. 2010. "How Are Habits Formed: Modelling Habit Formation in the Real World." *European Journal of Social Psychology* 40 (6): 998–1009. https://doi.org/10.1002/ejsp.674.

Lee, David S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *Review of Economic Studies* 76 (3): 1071–102.

Linden, Leigh L. 2008. "Complement or Substitute? The Effect of Technology on Student Achievement in India." *InfoDev Working Paper* No. 17.

Major, Louis, Gill A. Francis, and Maria Tsapali. 2021. "The Effectiveness of Technology-supported Personalised Learning in Low- and Middle-income Countries: A Meta-analysis." *British Journal of Educational Technology* 52 (5): 1935–64. https://doi.org/10.1111/bjet.13116.

Mandouit, Luke, and John Hattie. 2023. "Revisiting 'The Power of Feedback' from the Perspective of the Learner." *Learning and Instruction* 84: 101718. https://doi.org/10.1016/j.learninstruc.2022.101718.

McEwan, Patrick. J. 2015. "Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments." *Review of Educational Research* 85 (3): 353–94. https://doi.org/10.3102/0034654314553127.

Mo, D., L. Zhang, J. Wang, et al. 2015. "Persistence of Learning Gains from Computer Assisted Learning: Experimental Evidence from China." *Journal of Computer Assisted Learning* 31 (6): 562–81. https://doi.org/10.1111/jcal.12106.

Mojgani, Rebecca Sayre, Abbie Raikes, Jem Alvarenga Lima, Moth Pritchard, and Susan Graham. 2024. *From Feedback to Action: BEQI Feedback & Nudge Study in Mombasa, Kenya*. ECD Measure.

Moscoviz, Laura, and David K. Evans. 2022. "Learning Loss and Student Dropouts during the COVID-19 Pandemic: A Review of the Evidence Two Years after Schools Shut Down." *Center for Global Development Working Paper* No. 609.

Muralidharan, Karthik, Abhijeet Singh, and Alejandro J. Ganimian. 2019. "Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India." *American Economic Review* 109 (4): 1426–60. https://doi.org/10.1257/aer.20171112.

Naik, Gopal, Chetan Chitre, Manaswini Bhalla, and Jothsna Rajan. 2020. "Impact of Use of Technology on Student Learning Outcomes: Evidence from a Large-Scale Experiment in India." *World Development* 127: 104736. https://doi.org/10.1016/j.worlddev.2019.104736.

OECD. 2023. *PISA 2022 Results (Volume I): The State of Learning and Equity in Education*. OECD. https://doi.org/10.1787/53f23881-en.

Okonkwo, Chinedu Wilfred, and Abejide Ade-Ibijola. 2021. "Chatbots Applications in Education: A Systematic Review." *Computers and Education: Artificial Intelligence* 2: 100033. https://doi.org/10.1016/j.caeai.2021.100033.

Rodriguez-Segura, Daniel. 2022. "EdTech in Developing Countries: A Review of the Evidence." *The World Bank Research Observer* 37 (2): 171–203. https://doi.org/10.1093/wbro/lkab011.

Ryan, Richard M., and Edward L. Deci. 2000. "Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being." *American Psychologist* 55 (1): 68–78. https://doi.org/10.1037/0003-066X.55.1.68.

Semenova, Vira. 2025. "Generalized Lee Bounds." *Journal of Econometrics* 251: 106055. https://doi.org/10.1016/j.jeconom.2025.106055.

Sinclair, H. Colleen, Stacey R. Terrio, and Youn Kyoung Kim. 2025. *The Math Mind Measures: A User Guide*. Social Research and Evaluation Center at the Louisiana State University College of Human Sciences and Education.

Soe, Kyaw, Stan Koki, and Juvenna M. Chang. 2000. *Effect of Computer-Assisted Instruction (CAI) on Reading Achievement: A Meta-Analysis*. Pacific Resources for Education and Learning.

Tzenios, Nikolaos. 2020. "Examining the Impact of EdTech Integration on Academic Performance Using Random Forest Regression." *Researchberg Review of Science and Technology* 3 (1): 94–106.

Vegas, Emiliana, Lauren Ziegler, and Nicolas Zerbino. 2019. *How Ed-Tech Can Help Leapfrog Progress in Education*. Center for Universal Education at Brookings.

Verhoeven, Ludo, Marinus Voeten, Ellie van Setten, and Eliane Segers. 2020. "Computer-Supported Early Literacy Intervention Effects in Preschool and Kindergarten: A Meta-Analysis." *Education Research Review* 30 (100325). https://doi.org/10.1016/j.edurev.2020.100325.

Vygotsky, L. S. 1978. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.

Wenglinsky, Harold. 1998. *Does It Compute? The Relationship between Education Technology and Student Achievement in Mathematics*. Educational Testing Service, Policy Information Center. https://eric.ed.gov/?id=ED425191.

Whitson, C. David, and Jodi Consoli. 2009. “Flow Theory and Student Engagement.” *Journal of Cross-Disciplinary Perspectives in Education*, No. 1, Vol. 2: 40–49.

World Bank. 2018. *Learning to Realize Education’s Promise*. World Bank. https://doi.org/10.1088/0960-1317/21/12/125014.

Yang, Yihua, Linxiu Zhang, Junxia Zeng, Xiaopeng Pang, Fang Lai, and Scott Rozelle. 2013. “Computers and the Academic Performance of Elementary School-Aged Girls in China’s Poor Communities.” *Computers & Education* 60 (1): 335–46. https://doi.org/10.1016/j.compedu.2012.08.011.

Yeager, David S., Paul Hanselman, Gregory M. Walton, et al. 2019. “A National Experiment Reveals Where a Growth Mindset Improves Achievement.” *Nature* 573 (7774): 364–69. https://doi.org/10.1038/s41586-019-1466-y.

Zhang, Ling, James D. Basham, and Sohyun Yang. 2020. “Understanding the Implementation of Personalized Learning: A Research Synthesis.” *Educational Research Review* 31: 100339. https://doi.org/10.1016/j.edurev.2020.100339.

6 Appendices

A. Theory of Change

Darsel's theory of change is shown in the diagram (Figure 1) below. The key assumption is that, in low-resource settings that have low-quality school-based teaching and learning, AI EdTech can be a valuable additional input to the production of learning (based on the literature on EdTech improvements to instruction (Rodriguez-Segura 2022), although impacts are heterogenous, hence there is some uncertainty around this assumption). In the diagram, see arrows from OUTPUTS to OUTPUT EFFECTS, and from OUTPUT EFFECTS to OUTCOMES. The hypothesis is that student use of Darsel directly leads to increased math proficiency and math-adjacent cognitive skills (e.g., quantitative reasoning); an assumption this evaluation will test.

The chatbot is potentially an effective solution for three reasons. First, its AI-powered adaptive learning algorithms ensure that students constantly see tailored content "at the right level." Teaching at the right level (TaRL) is established in the literature as an effective intervention for improving learning (Banerjee et al. 2016). Second, the chatbot provides instructional content (through hints and explanations) to address learning gaps and increase proficiency. Together, this allows the chatbot to identify and extend each student's 'zone of proximal development' (the boundary between tasks a student is able to do independently and tasks they can do only with support), which is where the literature has established teaching is most effectively targeted to produce new learning (Vygotsky 1978). Lastly, the chatbot's design incorporates insights from neuro-educational research, which emphasizes the importance of motivation, novelty, and variety for learning (Howard-Jones 2010). This is done by including motivational messages and other gamification elements (e.g., levels and weekly leaderboards) to make the experience joyful, maximizing usage (see the last 3 boxes under OUTPUT EFFECTS). These features also are expected to improve student motivation and self-confidence (under OUTCOMES). This outcome, along with the primary outcome of higher math proficiency, leads to meaningful long-term impacts along multiple dimensions, including academic attainment and career aspirations (IMPACTs).

This theory of change – by which the chatbot is particularly well suited to increase learning beyond simply getting students to spend more time on task – is grounded in established research in the psychology of education (Ryan and Deci 2000; Csikszentmihalyi 1990; Hattie and Timperley 2007; Lally et al. 2010; Bandura 1977). The central idea is that the chatbot can increase self-determination, self-efficacy, engagement and flow, all of which are important inputs into successful learning.

First, the chatbot is likely to foster intrinsic motivation and self-determination, particularly by providing autonomy and competence, two core components of Ryan and Deci's (2000) self-determination theory. This is because Darsel's technology allows students to choose what skills they want to work on, and when and how long to practice.

Second, the chatbot is well positioned to increase self-efficacy, defined as a student's belief in their ability to accomplish a task. Self-efficacy is fostered by mastery experience (Bandura 1977), which the technology provides in abundance through adaptive difficulty, visible progress, and scaffolding.

Third, theories of engagement and flow highlight the importance of an 'optimal match' between students' skills and the difficulty of the task. Flow (a theory developed by Csikszentmihalyi (1990)) in particular, is characterized by absorption in "a challenging activity that requires skills, merging of action and awareness, concentration on the task at hand, clear goals and feedback" (Whitson and Consoli 2009, 41). The chatbot can encourage and sustain this state by continuously adjusting task difficulty, providing clear goals and immediate feedback.

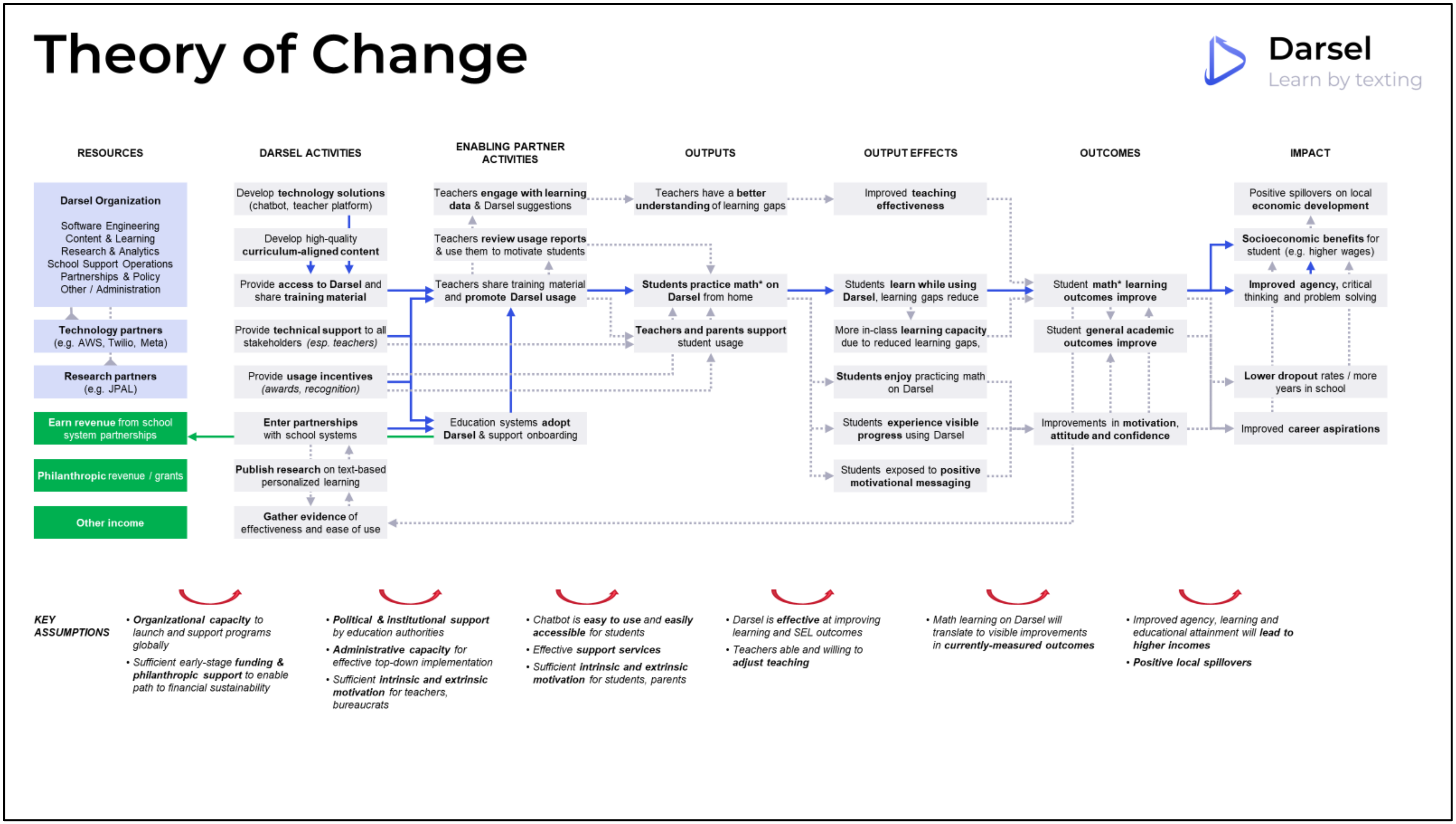
Fourth, the chatbot can provide immediate, task-oriented feedback as well as general encouragement, which can lead to positive emotions (see (Hattie and Timperley 2007) and (Mandouit and Hattie 2023)). Finally, nudges, reminders, and streaks can be used to build study habits (Lally et al. 2010). In all these dimensions, the chatbot can be used in ways that are not possible in a classroom with one teacher, many students and teacher-centered pedagogy.

Given higher usage of Darsel by female students (different OUTPUTS), impacts on areas such as career aspirations could vary by sex, particularly benefiting female students. All OUTPUTS (particularly usage), OUTCOMES, and near-term measures of IMPACT will be measured by the RCT to test this theory of change.

The intervention's broader theory of change assumes that behavioral change of various stakeholders is required to induce student usage of Darsel's product. Through its partnership with school systems, Darsel provides both support systems (e.g., training material, technical support through WhatsApp, with additional support for T2 teachers) and incentives (e.g., award ceremonies, recognition) which ultimately aim to maximize teachers' ability and willingness to support and encourage student use of Darsel (see ENABLING PARTNER ACTIVITIES). Teacher surveys and usage data will capture the extent to which these enabling partner activities take place.

The theory of change also recognizes the existence of several mechanisms that can increase impact. Students use Darsel from home, but this can affect in-class learning in several ways. First, chatbot-caused reductions in learning gaps could improve students' ability to benefit from classroom instruction (the third box under OUTPUT EFFECTS). Teachers may also adjust in-class instruction, due to the perceived effects of using the chatbot (T1 and T2) or in reaction to insights presented in Darsel's advanced reports (T2 only). Moreover, the socio-emotional learning benefits of using Darsel can raise learning outcomes (potentially beyond math) independently of direct proficiency gains from chatbot usage. Extrinsic motivation (e.g., student awards) may also contribute to improved learning outcomes.

Figure 1. Theory of change



Theory of Change

Darsel
Learn by texting

RESOURCES

Darsel Organization

Software Engineering
Content & Learning
Research & Analytics
School Support Operations
Partnerships & Policy
Other / Administration

Technology partners (e.g. AWS, Twilio, Meta)

Research partners (e.g. JPAL)

Earn revenue from school system partnerships

Philanthropic revenue / grants

Other income

DARSEL ACTIVITIES

Develop **technology solutions** (chatbot, teacher platform)

Develop high-quality **curriculum-aligned content**

Provide **access to Darsel** and share **training material**

Provide **technical support** to all stakeholders *(esp. teachers)*

Provide **usage incentives** *(awards, recognition)*

Enter partnerships with school systems

Publish research on text-based personalized learning

Gather evidence of effectiveness and ease of use

ENABLING PARTNER ACTIVITIES

Teachers **engage with learning data** & Darsel suggestions

Teachers **review usage reports** & use them to motivate students

Teachers share training material and **promote Darsel usage**

Education systems **adopt Darsel** & support onboarding

OUTPUTS

Teachers have a **better understanding** of learning gaps

Students practice math* on Darsel from home

Teachers and parents support student usage

OUTPUT EFFECTS

Improved **teaching effectiveness**

Students **learn while using Darsel**, learning gaps reduce

More in-class **learning capacity** due to reduced learning gaps,

Students enjoy practicing math on Darsel

Students **experience visible progress** using Darsel

Students exposed to **positive motivational messaging**

OUTCOMES

Student **math* learning outcomes improve**

Student **general academic outcomes improve**

Improvements in **motivation, attitude and confidence**

IMPACT

Positive spillovers on local **economic development**

Socioeconomic benefits for student (e.g. higher wages)

Improved agency, critical thinking and problem solving

Lower dropout rates / more years in school

Improved **career aspirations**

KEY ASSUMPTIONS

- ***Organizational capacity** to launch and support programs globally*
- *Sufficient early-stage **funding & philanthropic support** to enable path to financial sustainability*

- ***Political & institutional support** by education authorities*
- ***Administrative capacity** for effective top-down implementation*
- *Sufficient **intrinsic and extrinsic motivation** for teachers, bureaucrats*

- *Chatbot is **easy to use** and **easily accessible** for students*
- *Effective **support services***
- *Sufficient **intrinsic and extrinsic motivation** for students, parents*

- *Darsel is **effective** at improving learning and SEL outcomes*
- *Teachers able and willing to **adjust teaching***

- *Math learning on Darsel will translate to visible improvements in **currently-measured outcomes***

- *Improved agency, learning and educational attainment will **lead to higher incomes***
- ***Positive local spillovers***

B. Data Collection and Processing

The data will be collected through the following instruments and procedures:

- **Student Assessments and Surveys:**
    - Paper-based, proctored mathematics assessments and questionnaires administered in classroom settings.
    - Assessments and surveys will be proctored by trained staff to ensure standardization, compliance, and accurate data capture.
- **Teacher Surveys:**
    - Administered online via SurveyCTO, distributed through WhatsApp links provided to teachers by trained supervisors.
    - Follow-ups with non-responsive teachers will be conducted to maximize response rates.
- **Quality Assurance and Oversight:**
    - Supervisors and field operations staff from J-PAL will monitor proctors daily, conduct protocol checks, and report any irregularities.
    - J-PAL field operations staff and Jordan's Ministry of Education representatives will oversee implementation.
- **Logistics:**
    - Student assessments and survey sheets will be printed, securely transported, collected, scanned, labeled, and uploaded using encrypted digital storage.
    - All materials will be managed using a documented chain-of-custody to avoid data loss or mislabeling.
    - Disposal of physical documents will be secure and documented.

Administrative data on students' and teachers' engagement with the platform will be tracked during the intervention that will be implemented throughout the academic year (after baseline and randomization).

Data collection in each phase will conclude based on the following criteria:

- **Completion of Target Sample**: Data collection will stop once all participating students and teachers have been reached or all viable attempts have been exhausted.
- **Resource Constraints**: Data collection will be limited by time (estimated 10 working days per phase), available proctors/supervisors, and budget considerations.
- **Scheduled Duration**: Each data collection phase is expected to span no more than 2 weeks (estimated 10 working days).

Table 2. Grade 6 curriculum

#	Unit.Lesson	Unit Name	Lesson Name
1	1.1	1 Integer Operations	1 Integers and Absolute Values
2	1.2	1 Integer Operations	2 Comparing and Ordering Integers
3	1.3	1 Integer Operations	3 Integer Addition
4	1.4	1 Integer Operations	4 Integer Subtraction
5	1.5	1 Integer Operations	5 Integer Multiplication and Division
6	2.1	2 Fraction Operations	1 Fraction Addition and Subtraction
7	2.2	2 Fraction Operations	2 Mixed Fraction Addition and Subtraction
8	2.3	2 Fraction Operations	3 Mixed Fraction Multiplication
9	2.4	2 Fraction Operations	4 Fraction Division
10	2.5	2 Fraction Operations	5 Mixed Fraction Division
11	3.1	3 Decimal Operations	1 Decimal Multiplication
12	3.2	3 Decimal Operations	2 Decimal Division
13	3.3	3 Decimal Operations	3 Measurement and Units
14	3.4	3 Decimal Operations	4 Approximations
15	4.1	4 Translation and Geometric Constructions	1 Coordinate Planes
16	4.2	4 Translation and Geometric Constructions	2 Translation
17	4.3	4 Translation and Geometric Constructions	3 Reflection
18	4.4	4 Translation and Geometric Constructions	4 Circles
19	4.5	4 Translation and Geometric Constructions	5 Geometric Constructions
20	4.6	4 Translation and Geometric Constructions	6 Drawing Triangles
21	5.1	5 Algebraic Expressions and Equations	1 Powers and Exponents
22	5.2	5 Algebraic Expressions and Equations	2 Square root and Cube root
23	5.3	5 Algebraic Expressions and Equations	3 Order of Operations
24	5.4	5 Algebraic Expressions and Equations	4 Algebraic Expressions
25	5.5	5 Algebraic Expressions and Equations	5 Equations
26	5.6	5 Algebraic Expressions and Equations	6 Sequences
27	6.1	6 Ratios and Ratio Percentages	1 Ratio
28	6.2	6 Ratios and Ratio Percentages	2 Equivalent Ratios
29	6.3	6 Ratios and Ratio Percentages	3 Percentage and Fractions
30	6.4	6 Ratios and Ratio Percentages	4 Percentage and Decimals
31	6.5	6 Ratios and Ratio Percentages	5 The Percent of a Number
32	7.1	7 Geometry and Measurement	1 Quadrilaterals

#	Unit.Lesson	Unit Name	Lesson Name
33	7.2	7 Geometry and Measurement	2 Area of Parallelograms
34	7.3	7 Geometry and Measurement	3 Area of Triangle
35	7.4	7 Geometry and Measurement	4 Area of Trapezoid
36	7.5	7 Geometry and Measurement	5 The volume of a quadrilateral prism and its surface area
37	8.1	8 Statistics and Probability	1 Data collection
38	8.2	8 Statistics and Probability	2 Frequency Tables
39	8.3	8 Statistics and Probability	3 Frequency Tables and Frequency Diagrams
40	8.4	8 Statistics and Probability	4 Pie Charts
41	8.5	8 Statistics and Probability	5 Probabilities

Table 3. Math assessment structure by wave of data collection

	Baseline	**Midline**	**Endline**
	September 2025, before students start Grade 6	*Late November/ Early December 2025, after students learn Units 1-4 of Grade 6*	*May 2025, after students learn Units 5-8 of Grade 6*
Total number of items	29	29	30
Grade 6 unit-level prerequisite items			
Questions aligned to the Jordanian grade 4/5 curriculum, covering core prerequisite skills			
Unit 1 Prerequisite	2		
Unit 2 Prerequisite	2		
Unit 3 Prerequisite	2		
Unit 4 Prerequisite	2		
Unit 5 Prerequisite	2		
Unit 6 Prerequisite	2		
Unit 7 Prerequisite	2		
Unit 8 Prerequisite	2		
Grade 6 lesson-level items			
Questions aligned to the Jordanian grade 6 curriculum			
Unit 1 - Lesson 1	1	1	1
Unit 1 - Lesson 2		1	
Unit 1 - Lesson 3		1	
Unit 1 - Lesson 4		1	
Unit 1 - Lesson 5		1	
Unit 2 - Lesson 1	1	1	1
Unit 2 - Lesson 2		1	
Unit 2 - Lesson 3		1	
Unit 2 - Lesson 4		1	
Unit 2 - Lesson 5		1	
Unit 3 - Lesson 1	1	1	1
Unit 3 - Lesson 2		1	
Unit 3 - Lesson 3		1	
Unit 3 - Lesson 4		1	
Unit 4 - Lesson 1	1	1	1
Unit 4 - Lesson 2		1	
Unit 4 - Lesson 3		1	
Unit 4 - Lesson 4		1	
Unit 4 - Lesson 5		1	
Unit 4 - Lesson 6		1	
Unit 5 - Lesson 1	1	1	1
Unit 5 - Lesson 2			1
Unit 5 - Lesson 3			1
Unit 5 - Lesson 4			1

	Baseline	**Midline**	**Endline**
	September 2025, before students start Grade 6	*Late November/ Early December 2025, after students learn Units 1-4 of Grade 6*	*May 2025, after students learn Units 5-8 of Grade 6*
Unit 5 - Lesson 5			1
Unit 5 - Lesson 6			1
Unit 6 - Lesson 1	1	1	1
Unit 6 - Lesson 2			1
Unit 6 - Lesson 3			1
Unit 6 - Lesson 4			1
Unit 6 - Lesson 5			1
Unit 7 - Lesson 1	1	1	1
Unit 7 - Lesson 2			1
Unit 7 - Lesson 3			1
Unit 7 - Lesson 4			1
Unit 7 - Lesson 5			1
Unit 8 - Lesson 1	1	1	1
Unit 8 - Lesson 2			1
Unit 8 - Lesson 3			1
Unit 8 - Lesson 4			1
Unit 8 - Lesson 5			1
Global benchmark questions *based on TIMSS 2011*			
Topic 1	1	1	1
Topic 2	1	1	1
Topic 3	1	1	1
Topic 4	1	1	1
Topic 5	1	1	1

C. Cost-Effectiveness Analysis

The research will include a comprehensive cost-effectiveness analysis of Darsel, distinguishing between T1 and T2 in terms of both costs and impact. Darsel will share their detailed fixed costs (platform, grade, and country development costs) and variable costs (e.g., server/question costs) from their financial administrative data. Survey questions and administrative data on engagement with Darsel will allow for estimating time costs (for students and teachers), for both treatment arms.

The marginal cost per 0.1 standard deviation learning gain from the IRT scores for the mathematics assessment will be calculated for students as a key and (approximately) comparable summary measure. We will benchmark this against other EdTech and LMIC education interventions' cost-effectiveness (e.g., Muralidharan et al. 2019; Kremer et al. 2013; Angrist et al. 2025).

We will use the detailed cost-benefit analysis template constructed by The Abdul Latif Jameel Poverty Action Lab (J-PAL), available at https://www.povertyactionlab.org/media/file-research-resource/j-pal-costing-templatexls. We will follow the corresponding guidance provided at https://www.povertyactionlab.org/sites/default/files/research-resources/costing-guidelines_5.27.25.pdf.

The intervention expenses are primarily fixed/sunk costs from creating the Darsel platform, adaptive/AI software, and question bank development. Some costs recur for adding another grade's curriculum (e.g., adding grade 6) or adapting Darsel to the language and curriculum of another country. Marginal costs are quite low, estimated by Darsel at ~1 USD per student per year. This is notably half the cost of a highly effective intervention in India, which had a cost of ~2 USD per student per year at scale (Muralidharan et al. 2019). Darsel thus has the potential to be extremely cost-effective at scale.

D. Baseline Balance

We will test for baseline balance at the start of the academic school year for all outcomes (except drop out and grade repetition, due to their annual timing), along with all dimensions of heterogeneity. We will present differences for each variable between each pair of control, T1, and T2 (for both school-level treatment and classroom level treatment), along with the significance of those differences. We will also report F-tests for overall balance, based on a multinomial logit with clustered standard errors on the level of treatment (school or classroom), and its significance. We will include any imbalanced variables (5% significance cutoff) in the particular specification as controls, above and beyond the controls already specified.

7 Administrative Information

7.1 Funding

This research has received financial support from the Fund for Innovation in Development (FID), an independent initiative hosted by The Agence Française de Développement (AFD). Specifically, a Stage 2 Test and Position for Scale Grant was secured (project number 4672-JR).

7.2 Institutional Review Board (Ethics Approval)

The academic institution principal investigators (PIs) submitted a human subject research determination form to the University of Minnesota. Because the academic PIs are not involved directly in data collection nor will they have access to personally identifiable information, the university made the determination "Human subjects research, not involved" (STUDY00025886). J-PAL MENA undertook IRB review at the American University of Cairo (Case #2024-2025-289).

7.3 Declaration of Interest

Abdulhamid Haidar is both a co-author on this study as well as the CEO of Darsel. Darsel is the partnering organization for this study implementing the RCT alongside the government of Jordan's Ministry of Education. All other co-authors declare they have no conflicts of interest for the success of this study.

To address concerns about potential conflicts of interest, we note Darsel is a 501(c)(3) registered non-profit, and its purposes (per its official bylaws) include "conduct[ing] research in the educational field." While Abdulhamid does not stand to benefit financially from the research, he does have a vested interest in the results.

Several steps have been undertaken to minimize any risks of a conflict of interest affecting the study results, including:

- Development of a pre-analysis plan
- The questionnaires are being developed by other researchers (Abdulhamid and Darsel team did not write or select any of the specific questions)
- The math assessments will be developed by other researchers in direct collaboration with the Ministry of Education, and Abdulhamid (and the Darsel team) will not have access to the assessment questions
- Only the J-PAL team will have access to personally identifiable information
- Abdulhamid will review and can comment on analytical and data cleaning code, but will not generate code

- Data collection will be conducted by an independent data collection company that was selected by other researchers (Abdulhamid recused himself from evaluating bids), and they will provide data directly to the J-PAL team, without any role for Abdulhamid or anyone else in the Darsel team

7.4 Acknowledgments

J-PAL assisted in research management and the data collection process. MMIS Management Consultants was contracted to perform field management and data collection for the survey data collection. We thank Darsel for providing administrative data on the students using their application. We thank MoE for their collaboration in designing the assessment and implementing the study.

We would like to acknowledge our amazing research staff that is supporting the data collection and field implementation for this study. At J-PAL, Sarah ElBarky, Noha Adel, Nada Alaa, and Hady Aly provided research leadership and management. Nayera Adly Husseiny and Arwa Adel provided support throughout the initial design and fundraising process, as well as in policy communications. At Darsel, Mohamad Kan'an, Osama Zaidiah and Sarah Ayyash provided support related to collecting administrative data from the Ministry of Education and securing the necessary research permissions.

We appreciate Abbie Raikes and Rebecca Sayre Mojgani's generosity in sharing WhatsApp messages they used with teachers in a teaching quality feedback and nudge study (Mojgani et al. 2024) as examples for designing T2. We are grateful to Kyla Haimovitz for her assistance in identifying appropriate socio-emotional items for grade 6 mathematics students. We thank Giuliana Ghia for her contributions to the conceptualization and grant proposal, as well as the anonymous reviewers of the grant proposal who provided valuable feedback.