Scaling Down Costs to Scale Up Effective Programs: Evidence from Three Randomized Interventions

Sabrin Beg* Stephanie Bonds† Anne Fitzpatrick‡ Adrienne M. Lucas§

September 11, 2026

Abstract

To scale education programs, governments confront a critical trade-off between mitigating costs and preserving high-fidelity implementation. Teacher refresher training is often key to program fidelity, but delivering it in person at scale can be prohibitively expensive. Partnering with the Ghanaian Ministry of Education, we study this trade-off through three randomized controlled trials embedded in the nationwide roll-out of Differentiated Learning (DL) across 10,000 government primary schools. Comparing in-person to digital refresher training, we find that in-person refresher training significantly improved teacher implementation of DL by 7 percentage points and raised student test scores by 0.07 standard deviations, while digital refresher training had no effect. In a series of two subsequent experiments, we unpack the mechanisms underlying why digital training had no effect. We conclude that tactile, on-the-job training is needed to scale effective education programs; information on benefits to students and explicit directives from the Ministry of Education had no effect. These findings suggest that moving programs to scale using cheaper substitutes may not work because those substitutes lack the components that drive behavior change.

Keywords: professional development, on-the-job training, management, education, primary school, external validity, differentiated instruction

JEL Classifications: H40, I25, I28, M53, M54, O15, O43

*Department of Economics, University of Delaware. sbeg@udel.edu

†Department of Economics, Sociology, and Statistics, RAND Corporation. sbonds@rand.org

‡AEDE, The Ohio State University. Fitzpatrick.88@osu.edu

§Department of Economics, University of Delaware, 406 Purnell Hall, Newark, DE 19716. CGD, J-PAL, and NBER. alucas@udel.edu. 302.831.1901.

Acknowledgements: We gratefully acknowledge generous funding for the evaluation from USAID/DIV, FID, and UBS Foundation. Bonds acknowledges generous funding and time through a Post-Doctoral Research Fellowship at the Center for Global Development. For excellent research management, assistance, and policy-maker engagement, we thank Henry Atimone, Kanan Dubal, Sarah Kabay, Derrick Mensah, and the entire Innovations for Poverty Action Ghana team. For excellent research assistance we thank Raza Khan, Attique Ur Rehman, Manika Dhiman, and Mary Yilma. We thank Shahana Hirji for assistance with cost-effectiveness. This would not have been possible without our partners at the Ministry of Education, Ghana Education Services, the National Teaching Council, and UNICEF. This study is registered on the AEA RCT Registry: AEARCTR-0011414 and AEARCTR-0013844. The study was approved by the Innovations for Poverty Action IRB.

1 Introduction

Many interventions that produce large gains in tightly managed pilot settings fail to generate comparable impacts when scaled through routine public-sector systems (List, 2022; Vivalt, 2020). The scaling failure does not necessarily reflect flaws in the underlying intervention, but reflects the difficulty public bureaucracies face in supplying the complementary inputs required for high-fidelity implementation (Muralidharan and Niehaus, 2017; List, 2022).

In the education sector, which is the focus of this study, in-service teacher professional development is a common government tool to improve teaching. Yet, the versions implemented by governments at scale can be quite different from the successful versions implemented, and evaluated, at a smaller scale by non-governmental organizations or researchers (Popova et al., 2022). Programs that increase learning tend to include additional follow-up refresher trainings beyond the initial training but the marignal causal impact of these additional trainings on learning is unknown (Desimone, 2009). Governments can be hesitant to incorporate these refreshers because they are expensive, logistically complex, and difficult to sustain at a national scale. Further, refresher training may be seen as superfluous once the new methods become the status quo, making additional trainings inessential. Yet, refresher trainings themselves are more than a repetition of known information. They are a bundle of multiple, complementary inputs: reinforcement of technical knowledge, opportunities for simulated practice and peer exchange, and salient institutional signals that the reform remains a system priority. Two fundamental questions in scaling effective programs are 1) Are refresher trainings necessary to maintain fidelity of implementation at scale? and 2) Could some of the less expensive aspects of the full refresher training bundle be mimicked separately as a cost-effective way to maintain fidelity of implementation?

This paper presents experimental evidence on these questions in the context of Ghana's national scale-up of Differentiated Learning (DL) to 10,000 government primary schools. DL, where students are grouped at their learning level instead of grade level for part of the school day, a model closely related to Pratham's Teaching at the Right Level (TaRL), improves learning when implemented with fidelity (Banerjee et al., 2017; Beg et al., 2023; Duflo et al., 2024; Angrist and Meager, 2023). In recognition of its success across countries, and in Ghana specifically, the Ghanaian government received a USD$336 million funding from the World Bank that included funding for the scale-up

of the DL model to 10,000 government primary schools. Policymakers had experimental evidence in support of a DL bundle that included costly refresher trainings, but the marginal effect of the refresher training and its contribution to the overall DL program effect is unknown.

In partnership with the Ministry of Education (MoE) and its subsidiaries, we implemented three randomized controlled trials (RCTs) after the initial DL teacher training to test the importance of refresher training for program implementation and student learning outcomes, and whether lower-cost alternatives could sustain implementation. Prior to the start of all three experiments, teachers had received an initial 3-day in-person training, which involved teachers traveling to a central location to be trained by government training officials, a similar model and training used in Beg et al. (2023). At the onset of the study, teachers in all schools were supposed to be implementing DL as part of normal school operations. Additional in-person refresher training for all 10,000 schools would cost about USD $8 million and the government was interested in whether the expenditure was necessary and cost-effective.

The first RCT tested two models of refresher training relative to each other and business as usual Control schools. In the first treatment, In-Person Training, teachers participated in an additional centralized training of three days that provided comprehensive refresher training of the core components of DL. At the same time, teachers in second training, Digital Training, received access to digital training modules that contained similar content to the in person refresher training. Teachers could access these digital materials anywhere and work through them at their own pace. To encourage take-up of the Digital Training, within this treatment randomly selected teachers and their head teachers (i.e., school principals) received text messages and phone calls reminding them to complete the training. Teachers in control schools received no refresher training.

The second RCT tested a low-cost information-only intervention. As with the first RCT, teachers in all schools had received the initial training and were supposed to be implementing DL. In randomly selected schools, a trained facilitator conducted a school-based information session that provided evidence on the positive effects of DL and encouraged a group discussion on the evidence on the effectiveness of DL in Ghana and the importance of implementing DL as a government priority. As part of this session, teachers received a colorful, one page brochure that included compelling info-graphics and teacher quotes highlighting the success of the previous implementation of DL in Ghana.

The third RCT tested whether SMS reminders to implement DL affected implementation. This RCT was designed based upon teacher-reported feedback of institutional delays and lack of clarity from the Ministry was a driver of low observed rates of implementation. Using the same set of schools as the second RCT in the following academic year, we randomly assigned schools to an SMS reminder from GES that specifically provided that other schools were doing DL, with encouragement for them to implement.

These three RCTs unbundle the key ways that refresher training reinforces initial training to encourage fidelity of implementation at scale. We develop a simple framework in which the implementation of DL depends on a bundle of complementary inputs that enable teachers to know, value, and be able to perform the task. We identify four such inputs: (i) technical knowledge and skills of how to conduct DL, (ii) knowledge of its benefits, (iii) peer or community support that sustains practice, and (iv) salient institutional signals that the task is a priority. The interventions in our first experiment vary the intensity (and cost) of this input bundle. The In-Person refresher training provides (i) direct instruction and opportunity to practice how to implement DL, (ii) modules on understanding its benefits to students in previous iterations, (iii) opportunities for collaborative practice and troubleshooting (peer spillovers), and (iv) a strong signal of this being an institutional priority through bringing teachers to a central location where they were trained by ministry staff and received continuous professional development (CPD) credit. The digital refresher training offers a weaker version of the same bundle—lessons on implementation and the previous successes of the program, content created by ministry staff, and the signal of institutional priority through the existence of the content and the same CPD credit as the in person version upoon completion. To complete the course, teachers were supposed to record themselves practicing their teaching, reinforcing the skills to conduct DL. To strengthen the signal of institutional priority, randomly selected teachers and head teachers also received text messages and phone calls. The Information Intervention in the second experiment is a signal-only treatment: it raised salience and communicated benefits but did not increase knowledge of how to implement the program. The SMS text intervention in the third experiment is also a signal-only treatment: it raised salience and communicated the importance of DL, but did not provide any additional training.

We test the effects of the interventions on DL implementation in all RCTs. Additionally, we test for effects on intermediate outcomes related to our framework of (i) teacher take-up of training (ii)

teachers' conceptual knowledge about DL (iii) teachers' reported ability or skill to implement DL (iv) degree to which teachers form a peer network or community to learn about DL, (v) teachers' perception about DL as an institutional priority and (vi) teachers perceptions of the benefits and effectiveness of DL. In the first RCT only, we additionally test the effect of the interventions on student test scores.

We first document that the implementation of DL decays the farther teachers are from their initial training. In the control group, where teachers received the initial training and no refresher training, head teacher and teacher-reported implementation rates fall by one third and teacher reported implementation by two thirds over the course of an academic year.

We find that In-Person refresher training was the only intervention that significantly increased implementation, either self-reported or observed. The In-Person refresher about doubled the likelihood students participated in leveling exams, the first step in the DL process, and tripled the likelihood that students were observed divided into learning levels, the classroom manifestation of DL. Digital refresher training changed none of these outcomes, perhaps because of the low take-up rate: only 11 percent of teachers joined the digital training course platform and 6 percent completed the course. The In-Person refresher was also the only intervention that increased student test scores—0.07 SD or about 15 percent of a year of learning in this context.

We examine the potential mechanisms through which the interventions could have increased the fidelity of implementation of DL. None of the interventions increased teachers' conceptual knowledge of DL or teacher perceptions about the value of DL. We similarly do not find any evidence that either of the refresher trainings fostered the creation of a peer support community for teachers.

The In-Person refresher training improved implementation and student test scores without improving conceptual knowledge of DL or perceptions about its value and benefits, two components of In-Person training for which Digital training or an Information Session could most readily be substitutes. The results are consistent with DL needing to be learned experientially: knowing the theoretical practices is insufficient; teachers need to be able to practice it under supported conditions. In-Person refresher training uniquely provided opportunities for communities of teachers to come together across schools for simulated practice, peer feedback, and hands-on modeling. The In-Person refresher was likely effective in building the capacity or task-specific skills that are more critical to a successful and sustained implementation; teachers in the In-Person refresher training

arm were more confident in their DL ability when asked about how easy and comfortable they were in implementing different aspects of DL. Digital training was both less expensive and less effective in providing the same inputs for DL implementation.

Beyond its obvious policy implications as 28 countries across two continents are in the process of scaling up DL, the contribution of this paper to the economics literature is three-fold. First, we provide experimental evidence on the importance of high quality refresher training to maintain fidelity of implementation when scaling effective programs. We find that even after something is decreed as the new status quo in the system and initial training occurs, the human capital necessary to change practices is task-specific and depreciates over time (Topel, 1991; Lazear, 2009). Even something that is touted as straightforward, as differentiated learning is, requires repeated training. Capability-building—not just signaling and information—is a necessary condition for sustained scale-up in weak-incentive public systems.

Second, we show that in person refresher training provided the required additional reinforcement to increase take-up. Even though the refresher training was conducted by the same personnel and covered the same topics as the initial training in more condensed way, the additional reinforcement increased fidelity of implementation. Trying to replicate pieces of the training separately, was ineffective. Our information session built on literature that has demonstrated that when people receive information on the returns to an activity, even one that involves substantial investment such as additional schooling, migration, or changing sexual partners, they change behavior in response (Jensen, 2010; Baseler, 2023; Dupas, 2011). Our information session provided information on both the tangible (increased learning) and psychic (other teachers like it) returns to DL. We demonstrate the pitfalls of lower-touch substitutes for in-person training—providing pieces of the high-touch in-person training, did not replicate its effects.

Third, we provide evidence that while technology is a tempting way to improve the efficient provision of public services in developing countries, many barriers remain. Digital materials can deliver consistent information at a low marginal cost and provide a weak signal on program continuity. In our setting, teachers did not overcome the engagement hurdle. The most common use of technology in schools has been direct student instruction (Escueta et al., 2017; Singh et al., 2025; Beg et al., 2019). Outside of schools, technology has also been used in the delivery of training and information in agricultural extension (Aker et al., 2016; Fabregas et al., 2019), health service

delivery (Free et al., 2013; Ibeneme et al., 2022), and improving bureaucratic and political processes (Fujiwara, 2015). A growing literature explores digital or hybrid professional development as a scalable alternative to in-person training (e.g., (Mehmood, 2023; Davies et al., 2024; Hennessy et al., 2022)). We offer insights to inform when low-touch digital solutions can replicate high-touch in-person models. Digital models may eventually achieve similar impacts, but only if engagement is high, and they reproduce the essential ingredients that drive behavioral change. Well-designed digital models could substitute for in-person training, but only if they meaningfully recreate peer interaction, practice opportunities, and feedback loops. In our trial, engagement with the digital training was low and interaction nearly absent, with corresponding null effects. Finally, from a cost-effectiveness perspective, the results imply that the cheaper interventions may be more cost-effective only if they can effectively deliver the bundle of inputs required to generate behavioral change. In our context,both the digital training and the information intervention were too low-contact to recreate the mechanisms that increase take-up after in-person training and respondents failed to increase implementation, illustrating a broader lesson for the economics of scaling: the cheaper intervention is not necessarily more cost-effective if it omits complementary inputs that are core to sustained behavior change (Bold et al., 2018; List, 2022).

Taken together, the findings highlight that achieving durable impact at scale requires more than simply "going digital" or "sending signals"—it requires building the conditions that make skilled practice executable in the everyday reality of government schools.

2 Background

2.1 Ghanaian Education

Ghana is a lower-middle-income country in West Africa. Similar to other countries in Africa, primary school enrollment has increased over the past 25 years. In 2022, 98% of age-eligible children were enrolled in primary school (World Bank). These quantitative gains in enrollment have not been matched by commensurate improvements in student learning outcomes, contributing directly to the pervasive global challenge known as the "learning crisis" (World Bank, 2018).

The primary cycle is structured into Lower Primary (P1-P3) and Upper Primary (P4-P6). Each class is taught by a grade-level, and not subject-specific, teacher. The national government

maintains full authority over teacher employment, licensure, and curriculum design. A defining feature of this structure is the limited operational autonomy afforded to school head teachers, i.e., school principals, who possess minimal ability to hire, fire, or otherwise manage personnel.

Prior to being posted to a school, most teachers obtain a teaching certificate and teaching license. Once teachers start their teaching careers, they have both limited requirements and limited opportunities to engage in professional development. Since 2017, primary schools were mandated to hold 2-hour Professional Learning Communities (PLCs) meetings each week, but in reality, they were rarely held or did not cover appropriate material (Dampson, 2021). Teachers are supposed to earn 18 Continuing Professional Development (CPD) points each year to maintain their teaching licenses. Teachers can get these points by attending trainings, buying points without attending trainings, and signing up for virtual feedback sessions for teaching. Training is still viewed at all levels—from the Ministry of Education to the teachers themselves—as a "one-off" requirement, with the need for additional refresher trainings seen as redundant.

2.2 Differentiated Learning

One approach to solving the learning crisis is differentiated learning (DL), known in other contexts as Teaching at the Right Level (TaRL), a teaching method that divides students by learning level instead of grade level. DL has been proven to be effective in numerous contexts through rigorous randomized evaluations in settings as varied as Indonesia, India, and Kenya, including two in Ghana with effect sizes ranging from 0.11-0.14 (Banerjee et al., 2007, 2017; Duflo et al., 2011, 2024; Beg et al., 2023).

In the current implementation of DL, existing classroom teachers divide students by learning level across primary grade levels for part of the school day, four out of five school days per week. This is the model implemented as the "partial day tracking" model in Duflo et al. (2024) and Beg et al. (2023), also in Ghanaian schools. Teachers were to test their students at the start of each of the three school terms and divide their students by learning level instead of grade level for part of the school day. Students in P2 and P3 were divided among three learning levels, and students in P4 through P6 were divided into an additional three levels. To divide students among three levels in the lower primary grades, a head teacher or other spare teacher in the school would cover either

a learning level or the P1 classroom during DL time.[1] No additional teaching or support staff were added to the schools.

Based upon the evaluations of previous DL implementations, the Ghanaian Ministry of Education committed to scaling up DL to 10,000 of the 18,000 government primary schools as part of GALOP, the Ghana Accountability for Learning Outcomes Project.[2] As part of the scale-up, each district designated someone at the district level to be the District DL Coordinator. This person was supposed to ensure DL implementation within each district.

This study occurs after the initial DL in-person training for all GALOP schools. As part of the continuing implementation, the government wanted to understand whether additional, expensive refresher trainings were necessary.

3 Intervention Details

We partnered with Ghana Education Services to identify approaches to maintain DL implementation fidelity at scale. The interventions occurred across two similar, but distinct samples, of primary schools. All study schools received the initial DL training for teachers in 2021. The interventions took place in P2 through P6. See Section 5 for sample details.

Refresher Training Study

In-Person Refresher Training: The in-person refresher training was delivered in May 2023 and occurred over a period of 3 days, conducted by the same district level officials who conducted the original DL training. Each treatment school was invited to send five teachers to the training, ideally one teacher from each grade P2 through P6.[3] Head teachers (HTs) and school improvement support officers (SISOs), overseers responsible for approximately 8 geographically proximate schools, associated with these schools were also invited to the training. Teachers who completed this training earned 2 of the 18 annual required CPD points.

[1] Students in P1 were kept in their grade level as their learning levels are relatively more homogeneous than in other grades.

[2] GALOP was supported in part by the World Bank, Global Partnership for Education, and other donors. The project initially received funding of $218 million from the World Bank with subsequent funding received to total $336 million (World Bank, 2019). The 10,000 schools were selected because of their historical low performance, while also broadly ensuring schools across all districts received the program.

[3] Schools with multiple streams or sections of each grade could have sent multiple teachers from the same grade.

Digital Refresher Training: The Digital Refresher training covered the same content as in the in-person training and consisted of DL refresher modules delivered on a Digital platform.[4] Teachers in Digital Training schools were invited to a webinar orientation in June 2023, were registered on the platform, and given 3 weeks to complete all 5 modules, which had to be done in sequence. Each module contained lessons, assignments, and knowledge checks. At the end of the sessions, teachers were supposed to film themselves implementing DL and submit it through the portal for feedback and assessment. Teachers could access the course content through a WiFi connection or by using their own data bundles, which were not reimbursed. Teachers who completed the post-course survey that signified completion of the course received 2 of the 18 annual required Continuing Professional Development (CPD) points.[5]

Encouragement Nudges: During the three weeks in which teachers had access to the Digital portal, we randomly selected some teachers and head teachers to receive additional encouragement to complete the digital training. Within the Digital Training arm, teachers were randomized at the teacher level to receive (1) a follow-up SMS, (2) a follow-up SMS and a phone call, or (3) no additional encouragement. The SMS was tailored to the teacher's level of engagement with the portal. Those who had not started any lessons received a message encouraging them to start the materials. Those who had started but not finished were encouraged to complete the materials. The phone call was with a live person, started with a scripted message, similar to the SMS, based on a teacher's existing portal engagement with encouragement to start or complete the course. The conversation continued, providing instructions on how to access the course, reset the password, and download the materials to access offline at a later time if teachers were interested. Teachers were also reminded that they would receive 2 CPD points if they completed the course. This intervention was designed to be light touch and something that could be easily (SMS) or somewhat easily (SMS plus phone call) implemented at scale. Head teachers in the Digital Training arm were randomized at the school level to receive either a (1) phone call or (2) no additional encouragement. The phone

[4]The digital platform was accessible to all teachers, however the DL refresher modules were only intended for teachers in the digital refresher training arm. There was substantial effort towards ensuring that the two trainings were equal in content. The digital training was organized and administered by the National Teaching Council (NTC), who had previous experience implementing zoom trainings and webinars. In preparation for this digital training, representatives from GES and NTC met to agree on content, logistics, modality questions, and length.

[5]The data costs were somewhat offset as teachers did not have to pay the typical program cost for 2 CPD points of about GHS20-30 (approximately USD1.77 to USD2.65 in June 2023), and annually teachers have 1200 GHS per year for CPD points.

calls were similar to the teacher calls plus encouraging head teachers to support their teachers' completion of the modules.

Information Session Study

The Information Session treatment schools received a one page, double sided, informational pamphlet containing info-graphics about the success of DL in Ghana and positive quotes from interviews previously conducted with Ghanaian teachers with their thoughts about DL, e.g., "If you continue, it will change lives." The pamphlet was designed to be engaging for teachers. After the conclusion of all of the surveys at the school, enumerators held a structured group session to guide teachers and HTs through the contents of the pamphlets. This session provided an opportunity for teachers to ask questions and receive a deeper understanding of DL's goals, its alignment with the goals of the education sector, and the data presented in the informational materials. The aim was to ensure clarity and foster a shared understanding of DL's potential benefits and practical application in the classroom.

This intervention could have increased demand from teachers for additional conversations or training from school personnel about DL, reminded teachers of the intervention, and encouraged reticent teachers to implement the program.

SMS Text Study

Schools assigned to the SMS text message study received a series of weekly text messages from GES informing them that they were supposed to be implementing DL as part of GALOP, and reminding them to begin implementing DL in the classrooms that week. At treatment schools, the SMS messages were sent to all teachers and head teachers. In total, a series of 5 messages were sent to schools. This intervention was designed to address responses from teachers that they were not clearly given a dictate to implement DL, and convey to teachers that DL was part of the status quo.

4 Empirical Strategy

As each experiment use separate samples, we estimate the results separately. For the In-Person vs. Digital Refresher training, we estimate

$$y_{is} = \beta_1 \text{InPerson}_s + \beta_2 \text{Digital}_s + \Gamma' X_{is} + \varepsilon_{is} \tag{1}$$

where y_{is} is outcome y for respondent i in school s, InPerson_s is an indicator variable equal to one if the school was in the In-Person Refresher treatment, Digital_s is an indicator equal to one for schools in the Digital Refresher treatment, X_{is} is a vector of school and individual level controls (see below) including strata (district) fixed effects and baseline level of y_{is} as appropriate, and ε_{is} is a cluster-robust error term assumed to be uncorrelated between schools but allowed to be correlated within a school.

In student-level regressions, to improve precision, we include in X_{is} student age and age-squared, baseline grade level, an indicator variable for being female, and baseline test score for test score regressions. In teacher-level regressions, we additionally include in X_{is} teacher age, age-squared, years of experience, years of experience-squared, gender, and baseline class. In specifications with multiple waves, we include an indicator for the survey round in X_{is}. Regressions with outcomes on classroom-level implementation include in X_{is} the average percent of teachers present during the baseline survey.

The estimates of β_1 and β_2 provide the effect of each intervention relative to the status quo. The difference between the estimates of β_1 and β_2 provides the differential effect of In-Person vs. Digital Refresher training.

Within the Digital intervention, we estimate the effect of the various encouragement mechanisms as

$$y_{is} = \gamma_1 \text{SMS}_{is} + \gamma_2 \text{SMSCall}_{is} + \gamma_3 \text{HTCall}_s + \Gamma' X_{is} + \varepsilon_{is} \tag{2}$$

where SMS_{is} is an indicator variable equal to one if teacher i in school s only received an SMS, SMSCall_{is} equals one if the teacher received both an SMS and a phone call, and HTCall_s equals one if the head teacher of the school received a phone call. The first two are mutually exclusive and the third was cross randomized. X_{is} is a vector of control characteristics, including gender of

teacher and a fixed effect for our randomization strata (region).

For the information intervention, we estimate

$$y_{is} = \alpha \, \text{Information}_s + \Gamma' X_{is} + \varepsilon_{is} \tag{3}$$

where Information_s is an indicator variable equal to one if the school s received the Information session intervention. The controls are the same as in the other regressions.

For the SMS Text intervention, we estimate

$$y_{is} = \lambda \, \text{SMS Text}_s + \Gamma' X_{is} + \varepsilon_{is} \tag{4}$$

where SMS Text_s is an indicator variable equal to one if the school s received the SMS Text session intervention. The controls are the same as in the other regressions.

5 Sample Selection and Data

5.1 Refresher Training Study

The study was carried out in 230 schools in 46 districts, which our partner UNICEF selected because they had existing relationships in these districts. All schools received the initial DL training prior to the study. The 230 schools were randomly divided into In-Person Refresher training (50 schools), Digital Refresher Training (90 schools), and control (90 schools). We additionally conducted an encouragement intervention within the Digital Training arm, where teachers that were registered for the digital training platform by GES were randomized into the treatment arms; one third of the teacher sample was randomized at the teacher level into each of SMS, SMS plus phone call, and no encouragement and one half of the school sample was randomized at the school level into head teacher phone call (45 schools) and nothing (45 schools).[6] Figure 1 Panel A contains the randomization design for the refresher training study.

[Figure 1 about here]

Baseline Data: All study schools had received the initial DL training in March 2021. The

[6]Between baseline and follow-up, one school was dropped because it closed and five schools were dropped because of concerns for enumerator safety due to nearby insecurity and violence.

baseline occurred in September 2022, prior to the end of the 2022 academic year. We spoke with head teachers, teachers, and students, invigilated student exams, and conducted classroom observations. Within each study school, the head teacher, the teachers of grades P2-P6, and four randomly selected students present on the day of enumeration from each grade P2-P6 (16 total students per school) were selected as respondents. If schools had fewer than four students in a grade present on the day of enumeration, all present students were selected. Appendix Table A1 contains the baseline balance. We do not find any statistically significant differences across the three treatment arms. Self-reported measures of DL indicate that at least initially the scale-up of DL was moderately successful: 63% of head teachers report that they were implementing DL. Teachers similarly report high rates of implementation at 57%.

Follow-up Data: The follow-up was unannounced and occurred August 1 to September 3, 2023. As with the baseline, it included surveys of all of the respondents, an exam for the same students from baseline, and classroom observations.

Digital Portal Data: For each teacher in our sample in the Digital Training arm we have data from the digital portal, which indicates whether or not a teacher completed a series of checkpoints necessary to complete the course. These include whether or not they joined the WhatsApp group associated with the course, completed the pre-course survey, completed each module of the course, and completed the post-course survey. We further have the teacher scores from the 12 graded quizzes throughout the course that teachers had to complete, but did not have to pass, to continue to the next part of the course.

5.2 Information Session Study

The study occurred in 200 schools across all 16 regions of Ghana selected from the list of 10,000 GALOP schools. The sample was randomly divided into Information treatment (100 schools) and control (100 schools). All schools received the initial DL training prior to the study, and all schools in this study were supposed to be implementing DL as part of the status quo.

Baseline Data: We conducted the baseline in February 2024. The Head teachers and five teachers, one from each grade level P2 through P6, were surveyed at each school. Table A2 shows the baseline teacher and school characteristics in our study sample. This baseline occurred in the school year after the Refresher Training baseline. Consistent with the increased time elapsed since

the initial teacher training on DL, only 35% of head teachers reported their school was implementing DL at baseline, lower than the self-report in the Refresher Training baseline. Otherwise these schools are similar to those in the other intervention.

Post Session Data: Immediately after the pamphlet distribution and information session, enumerators surveyed teachers who had participated in the information session asking questions about participants' comprehension of the information provided.

Follow-up Data: In the follow-up survey, the same schools as the baseline were visited between June 3 and July 15, 2024, the third term of the academic year. Enumerator teams surveyed head teachers and five teachers from each school, attempting to reach the same teachers even if the teacher was teaching a different grade level in the school. They also conducted classroom observations. Schools were aware that a research team would be visiting them but the exact date of the visit was unannounced.

[Figure 2 about here]

5.3 SMS Text Study

This study occurred in the same 200 schools as the Information Session, during the subsequent academic school year. The sample was randomly divided into SMS Text treatment (100 schools) and control (100 schools), stratified by region and Information Session treatment status. Similar to other studies, all schools in this study were supposed to be implementing DL as part of the status quo. Table A3 demonstrates that the randomization was successful at creating comparable groups.

Follow-up 2 Data: In the follow-up survey, the schools were revisited during the third term of the academic year. The same protocol was followed as in the Follow-up round of the Information Session study. Enumerator teams surveyed head teachers and five teachers from each school, attempting to reach the same teachers even if the teacher was teaching a different grade level in the school. They also conducted classroom observations. Schools were aware that a research team would be visiting them but the exact date of the visit was unannounced.

Figure 2 shows the timeline for the the studies and important school year milestones.

6 Results

6.1 Take-up and Engagement

Before examining the change in implementation fidelity due to the treatments, we confirmed that the refresher trainings occurred as intended (see Table A4). Among teachers in the In-Person refresher group, 86% reported attending. The teachers in the Digital refresher arm did not have a higher likelihood of attending an off-site training.[7] We test if the Digital refresher training increased engagement with the Digital Portal available to all teachers and where the Digital refresher training modules were posted. The teachers in the Digital Training were 11 percentage points more likely to report signing up on the Digital Portal relative to the control group.[8] We corroborate this self-reported engagement with back-end data from the Portal. Nine percent of teachers in the Digital refresher completed the pre-course survey, consistent with the 11 percent difference found in the table (see Appendix Table A5).

This low engagement with the Digital Portal by the Digital Training group occurred even with additional reminders and nudging as part of the Encouragement sub-RCT (Appendix Table A5). Regardless of the nudge, teachers were no more likely to join the course WhatsApp group (column 1) or complete the pre-course survey (column 2). Teachers who received both an SMS and a phone call were 9.2 percentage points more likely to complete the post-course survey, the requirement to earn CPD points (statistically significant at the 10 percent level). In sum, none of the nudges appeared to meaningfully increase rates of engagement with the digital platform. At a maximum, with the most effective nudges, completion of the digital training course increased to only 11 percent.

For the Information Session sample, we measured beliefs immediately after the information session (Table A6). We find that teachers are more likely to agree that DL is effective compared to traditional learning immediately after the invention, however, there is no persistent change in

[7] Recall that DL should have been in all schools as the status quo——32 percent of teachers in control schools reported being trained on DL in the past year at a central location. The initial training was more than one year prior to our follow-up survey. Control teachers may have incorrectly remember, conflated this training with another one, or the SISO or other district-level officer could have conducted training for the schools, which they were encouraged to do as part of normal DL operations. The In-Person Refresher group is both statistically different from the control group and the Digital intervention.

[8] Forty-eight percent of control teachers reported signing up. Control teachers could have confused this with the DL Teacher Telegram App that contained digital copies of DL materials and the DL teacher's guide, to which all teachers had access, or occasional webinars that GES and NTC hosted. Teachers outside the Digital intervention could access the refresher training modules, but only the digital treatment teachers were informed about them during the webinar.

beliefs due to the intervention. During the follow-up rounds, the beliefs about DL are equal in the treatment and control groups.

6.2 DL Implementation

We turn to whether any of these interventions improved DL implementation. Recall that *all* study schools had received initial training in and were supposed to be implementing DL, even the control schools. DL consists of first conducting a leveling exam for all students. Subsequently, teachers must split students into those levels for 4 out of 5 days of the week.

Results on DL implementation are presented in Table 1. We examine various teacher reported measures of different aspects of DL implementation. In the Refresher Training Sample (Panel A), control group teachers report that 28% of their students participated in the leveling exams. Both refresher training interventions increase the likelihood that the students participated in leveling exams by 23 percentage points in the In-Person arm and 9.8 percentage points in the Digital Training arm (Column 1). However, our data suggest that this first step only translates into partial implementation. While teachers in the In-Person arm report higher likelihood of splitting students at least once in the past 5 days, there is no significant change for teachers in the Digital group (Column 2). Similarly, teachers in the In-Person group report an increase of 0.59 days of splitting students into levels, from a base of 0.7 days per week, with no significant changes in the Digital training arm (Column 3). In the control group, only 18 percent of schools reported that they were implementing DL on the day of the spot check; this rate is higher for teachers in the In-person and Digital training arms, and statistically significant for the Digital arm at 10% level (Column 4). These results differ from enumerator-reports of this unannounced spot check (Column 5). Enumerators report that only 3% of schools in the Control arm are implementing DL. This rate meaningfully increases by 7 percentage points for schools in the In-Person training arm, but is not different for schools in the Digital Training arm. Head teachers' reported DL implementation rate is also higher in the in-person training schools (Table A7).[9]

In A1, we graph the change between baseline and endline separately by group using an index of

[9]Since we test multiple outcomes as measures of Implementation, we test the robustness of our findings after correcting for multiple hypothesis testing using the Anderson (2008) method to control for the false discovery rate. The adjusted q-values in Table A8 show that the effects on implementation are significant as before, except the coefficient on observer reported implementation.

DL implementation, using a combination of teacher-reported values for implementation and doing the leveling exam. In the control group there is a sharp decline in implementation over time; the in-person training reverses that trend. Digital training slightly attenuates the overall decline that would have happened in the status quo. Overall, the implementation rate in the control group is low—only 18% of control teachers report doing DL that day (compared to 57% at baseline) and only 3% of classes are organized by student level according to observer reports. The In-Person training is the most effective in increasing the likelihood that teachers say they are doing DL.

The results in Panel B similarly show that the Information Session did not significantly change implementation rates. In this sample, 18% of teachers report completing the leveling exam and 22% report implementing DL on the day of the survey, while enumerators' report 11% of schools split students by levels (Column 5). The Information session did not change DL implementation by any of our measures.

The results in Panel C reinforce that the low-cost approaches to nudging teachers to implement DL in the classroom were ineffective. The SMS text mesasge did not change DL implementation by any of our measures. However, there is a general boost in implementation for the entire sample. In this round, 29% of teachers report completing the leveling exam and 27% report implementing DL on the day of the survey. However, enumerators' still report 11% of schools split students by levels (Column 5).

6.3 Student Attendance and Learning

In Table 2 we examine whether DL implementation translated into improved student learning. Recall that we only conducted student tests in the Refresher Training experiment. We find that learning increases with refresher training. Test scores for students in In-Person Training schools increase by 0.07 standard deviations (Column 1), with significant gains in English (Column 2). Math scores are larger, but not precisely estimated (Column 3).

In Table A9, we test for attrition by treatment status. We find no effect on student attendance (mean student attendance is relatively low (64%), and this rate is balanced by treatment group). However, with additional tracking we were able to conduct assessments with 94% of students, a rate that is 3 percentage points lower in the digital refresher training group (Column 2). Thus, we also present Lee Bounds in Appendix Table A9 (Lee, 2009) and rule out differential attrition by

baseline test score (Appendix Table A10).

In Appendix Table A11, we also test if the intervention affected teacher attendance and find some evidence that the In-Person refresher training intervention increased the retention rate for teachers, i.e., teachers are more likely to be at the same school at the followup. This may be because the teachers were participating more actively with the DL scale up, lowering turnover and transfers. We do not find any treatment effect on the likelihood that the teacher is present at school.

6.4 Mechanisms

In Table 3, we examine barriers to implementing DL, to advance the science of scaling, focusing on complementary inputs required for successful implementation at scale that may be addressed through refresher trainings. We examine (i) knowledge about the DL task, (ii) perceptions about DL as an institutional priority, (iii) the degree of peer support for DL, (iv) perception about the benefits of DL and (v) teachers' reported ability to conduct the various aspects of DL. We find that Control group teachers in both experimental samples score 55% on a set of questions about the DL task. None of the interventions increase knowledge about DL but the In-Person intervention did increase implementation, implying that technical knowledge is not the primary barrier to scaled implementation in this context as the interventions increased. Similarly, the low rates of implementation is not due to a lack of understanding of the importance: 75 to 85% of teachers report agreeing that the Ministry of Education cares if the teacher implements DL (Column 2). The refresher training interventions do not change perceptions about the importance of DL. Thus, we infer that the in-person refresher training is unlikely to have operated by signaling the Ministry's prioritization of DL.

Next, we examine if refresher trainings lead to additional peer engagement and support (Columns 3-5). In the existing model of continuous professional development, teachers in the status quo should be attending weekly PLC sessions, where peer-to-peer learning should occur. They are further monitored and encouraged to implement DL by their head teachers, SISOs, and the District DL Coordinators. Within the PLC framework, teachers are also able to record their teaching and get peer review and coaching from the virtual feedback sessions with senior teachers. Therefore, we also test whether any of the interventions increased peer learning and training at schools. We ask the teachers if they received DL instruction from their HT or from another teacher in their school

and if they received peer observation and feedback for their recorded teaching.

We find that the interventions spurred some additional training at schools. In the Digital Training arm, the likelihood of receiving DL instruction from other teachers is marginally higher for the treated teachers (Column 4). Teachers are also significantly more likely to receive one-on-one peer observation and feedback in the Digital arm (Column 5). Even though engagement with the Digital materials was low, the limited engagement, its existence, and reminders about its existence may have caused some peer interactions and higher coaching support for teachers. Since peer learning and support is unchanged in the in-person training arm, this channel is unlikely to drive higher implementation due to the in-person refresher training.

Finally, in Column 6 we test if the intervention moved the perceived benefits of DL. We observe that 9% of teachers prefer DL over the standard curriculum, and this rate is higher in the In-Person training arm. Thus, teachers in the In-Person training arm have higher preference for DL, which could be a possible channel for higher implementation, or conversely an effect of teachers implementing DL.

While the Information Session was not training, we similarly test for effects on knowledge, perceptions about importance or benefits of DL, and peer support. While we do not observe any differences in training on DL from head teachers, we do observe a 8 percentage point increase in the likelihood of receiving instruction on DL from another teacher. These results are consistent with school personnel providing additional training on DL in response to demand from teachers who were reminded about its importance and success through the Information Session. There are no effects on the other measured channels and complimentary inputs for implementation in the Information intervention. We similarly find no changes from any of the tested mechanisms in the SMS text message intervention.

In summary, In-Person refresher was effective in increasing implementation but not through testable knowledge, perceptions about importance of DL, or peer support. Instead, as In-Person training increased implementation while the other interventions did not, the In-Person refresher likely imparts DL related practical skills that are critical for implementation, ensuring that teachers were better equipped to implement DL. Teachers' self-reported ability to teach DL is higher as shown in Table 4. The teachers in the In-Person refresher training arm are more likely to report that they can teach DL, that they find conducting DL exams easy, and that they use DL materials

outside class.

We additionally examine teacher reported barriers for DL implementation in the refresher training sample and test if any of these constraints are relieved by the In-Person refresher. Figure 3 shows that the primary teacher-reported reasons for not implementing DL include planning and other delays, materials, and training/understanding of DL. Consistent with importance of in-person practice and skill development, the In-Person refresher reduced the barrier of not understanding how to implement DL or having inadequate training. In the control group, 15% of teachers report that their primary barrier to DL implementation is training or understanding. This is 8 percentage points lower among the In-Person refresher group, suggesting that the In-Person refresher improved teachers' understanding of how to implement DL.[10]. Thus, in summary, the in-person refresher training may have provided teachers the task-specific skills to implement DL.

7 Cost-Effectiveness

We use the ingredients method to assess cost effectiveness of each intervention in 2025 USD (Hirji et al., 2023).

In-Person Training. The total cost of the in-person refresher training was $40,486 or $128 per teacher across our 49 In-Person training schools including training expenses and the transportation of participants to these central trainings. This does not include the development of materials as materials had already been developed as part of the initial trainings. Enrollment in the In-Person schools in treatment grades was 6,395. Therefore the per student cost was $6.33. The In-Person training increased test scores by 0.071SD, or 15 percent of a year of learning based on the control group learning 0.48SD over the year. To aid in the comparison across studies, this is 1.12SD per USD$100 or 2.34 years of learning per USD$100. To scale another way, this is USD$42.80 per year of student learning or USD$89.17 per standard deviation increase in test scores. These calculations make assumptions about the linearity of returns relative to the dosage but are provided for the readers' convenience. When comparing across studies, recall that students were taught by trained teachers for only about five months prior to follow-up student assessment.

[10]Teaching and DL materials was another common barrier among the control group. While we do not find any significant difference at endline in teachers' reporting materials as a constraint, we separately asked teachers if they generally receive teaching materials at in-person trainings. Teachers in the in-person arm are more likely to report receiving materials at trainings (Appendix Table A12)

Digital Training. The Digital Training did not increase test scores. Nevertheless, we provide costs for the interested reader. The total costs of the digital training intervention were USD$4,622, USD$7.34 per teacher, or USD$0.40 per student. This includes the costs of hosting but not the development of the materials as the materials had already been developed as part of the initial training. The per student costs of the Digital Training were about 6 percent of the costs of the In-Person training. Therefore, even small test score increases would make this type of intervention more cost effective than In-Person training.

Within the Digital Training we provided additional nudges to some teachers and head teachers. The cost of the head teacher phone calls including developing the script, training the caller, the caller's time, and the caller's airtime was USD$13.61 per head teacher or USD$0.10 per student. The cost of the teacher SMS including developing the SMS scripts and airtime associated with sending the messages was USD$4.37 per teacher or USD$0.22 per student. Finally, the cost of the teacher SMS plus call, which incorporated all of the above aspects, was USD$9.13 per teacher or USD$0.44 per student, more expensive at a per student cost than the digital training intervention.

Information Session. We do not have test score outcomes for students in the Information Session intervention. As the session did not change teacher behavior in the classroom, it was unlikely to have changed student test scores. The total cost of the information session was $1,128, including opportunity costs of staff time to design and create the pamphlet and facilitator wage. At the scale implemented, the costs were USD$1.61 per teacher or USD$0.08 per student, about 80 percent of the cost of the calls to the head teachers on a per student basis.

8 Discussion

In this section we discuss our results in the context of a brief conceptual framework and assess external validity and generalizability.

8.1 Conceptual Framework and Discussion

To assist our discussion of results, we conceptualize the implementation of Differentiated Learning (DL) as a task that requires a bundle of inputs held by teachers. Let i index teachers and let the

probability that a teacher implements DL with fidelity, I_i, be denoted by

$$I_i = f(S_i, K_i, B_i, T_i, C_i),$$

where S_i represents the strength of *signals of institutional priority*, K_i denotes *knowledge of the task*, B_i captures *knowledge of the benefits* of DL, T_i reflects the teacher's *task-specific skills or capital* required for effective implementation, and C_i represents *community support and peer interactions*. The function $f(\cdot)$ is increasing in each of its arguments.

This framework nests several constraints that may bind when governments attempt to scale an intervention such as DL in resource-constrained public systems. First, teachers may lack knowledge about how to implement DL (K_i), or about its benefits for student learning (B_i). Second, weak monitoring and managerial infrastructure may imply that institutional priorities are unclear, reducing S_i and lowering the salience of DL. Third, teachers may lack opportunities to practice the task and thereby accumulate task-specific implementation skills (T_i). Finally, teachers may have insufficient supportive peer networks (C_i) that help sustain motivation and share tacit knowledge. The scaling problem thus arises when public systems attempt to expand DL without relieving these constraints.

Our experimental design maps directly onto this set of mechanisms. The In-Person refresher training, IP_i, is designed to shift multiple inputs simultaneously:

$$\text{IP}_i \rightarrow (S_i^{+}, K_i^{+}, B_i^{+}, T_i^{++}, C_i^{++}),$$

where the double positive superscript indicates a stronger expected effect. The training provides practice opportunities and hands-on demonstrations, enabling teachers to accumulate task-specific capital, and creates peer communities that may sustain support beyond the training event. The digital refresher training, DG_i, is expected to operate on the same mechanisms, though less intensively:

$$\text{DG}_i \rightarrow (S_i^{+}, K_i^{+}, B_i^{+}, T_i^{+}, C_i^{+}),$$

since engagement levels may be lower and peer interactions more limited. By contrast, the information session intervention and accompanying pamphlet, PM_i, primarily acts as an informational

signal and a message about DL's benefits:

$$\text{PM}_i \rightarrow (S_i^+, B_i^+).$$

Figure 4 shows the interventions and the inputs they supply.

Empirically, we find that none of the treatments generate detectable changes in either teachers' knowledge of how to implement DL (K_i) or their perceptions of the Ministry's prioritization of DL (S_i). Likewise, perceived benefits (B_i) do not shift differentially across arms. Nonetheless, both DL implementation and student learning improve substantially in the In-Person refresher training arm. In our framework, these results imply that the effects operate primarily through the accumulation of *task-specific capital* (T_i). The In-Person setting provided opportunities to rehearse DL techniques and build tacit skills that are not easily transmitted through digital materials or informational messages alone. This interpretation aligns with the literature showing that job- and task-specific human capital plays a central role in performance and can depreciate without active practice, e.g., Topel (1991); Ben-Porath (1967); Hanushek and Rivkin (2010). Our results also support this interpretation as the In-Person teachers report more confidence in conducting DL, and are less likely to report "Training and Understanding" as a barrier to implementation.

Digital refresher training failed to generate comparable impacts. This may occur if the Digital modality is ineffective in delivering the skills most essential for sustained implementation. Alternately, it could be that the digital training could provide the needed inputs but the teachers do not engage with it sufficiently. We test for heterogeneity in take-up of the Digital training. However, we do not find evidence that take up of the Digital training is associated with higher levels of implementation. In the Digital training arm, teachers with below median experience and teachers in schools with more amenities are more likely to engage with the Digital Platform (Table A13). However, these groups of teachers are not the ones who implement DL more. Table A14 shows that teachers with higher experience, higher baseline DL implementation and those in schools with below median amenities are more likely to implement DL. This tells us two things: first, the lack of treatment effect in Digital Training is not entirely because of low take-up and plausibly because the Digital refresher is unable to effectively provide the same complimentary inputs as In-Person refresher training. Second, digital modalities could thus still be leveraged to provide these inputs if

designed and implemented to transmit the required set of skills. Given the relatively lower costs of digital training, it has the potential to be a cost-effective alternative if teachers are able to engage sufficiently to learn and practice the required teaching skills.

Our findings highlight a key challenge for scaling: acquiring the necessary task-specific capital may require interactive, practice-oriented training environments. Cost-effective digital solutions may still be feasible, but only if they are designed to replicate the practice, feedback, and peer-interaction channels that are crucial for building the specific capital required for implementation.

8.2 External Validity and Generalizability

Because governments across low- and middle-income contexts struggle to bridge the gap between formal adoption of pedagogical reforms and frontline implementation, the mechanisms identified in this study—particularly the role of task-specific human capital and experiential training—have broad relevance. We assess external validity and generalizability along four standard dimensions: selection, attrition, naturalness, and scale.

Selection. The core implementation challenges we study—teachers being asked to adopt a new instructional practice, weak managerial incentives, and limited opportunities for hands-on practice—are common across education systems worldwide. DL or similar pedagogical routines are now embedded in national curricula in many low- and middle-income countries, yet actual take-up varies widely. The districts in our sample are broadly representative of the national system along observable characteristics, and the institutional environment—centralized hiring, limited performance incentives, and moderate administrative capacity—is typical of many public-sector settings. Relative to other governments, Ghana's Worldwide Governance Indicators rating is above the average for sub-Saharan Africa, but below that of India's and the worldwide median (World Bank 2018). Therefore, Ghana's level of state capacity does not stand out as unique but is similar to many countries. Thus, while the precise institutional details reflect the reality of our study country, the underlying mechanisms we test are relevant to a much larger set of systems where teachers face similar constraints on time, autonomy, and skill acquisition.

Moreover, the 'learning crisis' is not unique to Ghana or to Africa—it afflicts communities in some of the richest countries in the world. Many countries, regardless of their levels of income, have at least some locations where students are behind grade level (List 2022). Similarly, many

countries have public sectors with weak incentives. Our broad lessons are relevant to all of those countries. Our exact intervention might be most relevant to locations with similar characteristics, specifically the 53 countries with the same World Bank Lower-Middle Income categorization as Ghana (3 billion people), some of the richer countries in the 26 country Low Income category, some of the poorer countries in the 54 country Upper-Middle Income category, and the 47 countries of the sub-Saharan Africa region (1.2 billion people).

Attrition. Attrition was very low in our study. We were unable to follow-up in 6 study schools because of security concerns. Student attrition was minimal, with 94 percent of students participating in the achievement follow-up.

Naturalness. The tasks required of teachers, head teachers, and school support officers closely matched the activities they routinely engage in when new programs are introduced. In most systems, teachers attend in-service training when learning new pedagogical methods and receive communication from the school leaders or central education authorities. The interventions in this study therefore operate within normal administrative channels, without introducing artificial incentives or external implementers that might inflate effectiveness.

Scale of the Study. The design mirrors how a government would plausibly scale such reforms: all materials were produced by the ministry, trainers were existing government personnel, and all activities were embedded in standard workflows. Since 2022, the government has been replicating this program at scale in 10,000 government schools. The Digital Refresher and Information Sessions variants used communication channels the ministry already deploys for disseminating information. This implementation architecture provides a realistic test of the scalability of each approach, particularly the limits of low-cost digital refreshers when task-specific skills require experiential practice.

Given the representativeness of the sample, minimal attrition, the naturalistic embedding of tasks within routine government processes, and the scale-consistent implementation model, this study offers strong external validity for contexts in which workers must adopt tasks and programs under limited supervision and weak incentives. More broadly, the findings speak to public-sector scaling challenges where effective delivery hinges not only on knowledge or prioritization, but on the accumulation of task-specific capital that lighter-touch digital or informational interventions may fail to generate.

9 Conclusion

This paper evaluates what is needed to sustain the implementation of a proven pedagogical reform at scale. In partnership with the Ghanaian Ministry of Education, we conducted three randomized controlled trials during the national roll-out of Differentiated Learning (DL) to 10,000 schools. Across the studies, schools were randomized to receive In-Person refresher training or Digital refresher training; an Information Session; an SMS text message. Only the In-Person refresher training increased teacher implementation and improved student learning. Digital training was far less costly, but saw limited engagement, even when paired with SMS and call-based encouragement. The other low-cost approaches, such as the Information Sessions pamphlet or the SMS text message interventions changed teaching practices or student learning.

The results highlight a central constraint in scaling complex public-sector tasks: information and motivation are not substitutes for the capability required to execute skill-intensive practices. DL is a high-touch instructional method that relies on tacit, task-specific human capital built through demonstration, practice, and feedback. In-Person refresher training provided these experiential components. Digital and signal-only models did not. That these lighter-touch interventions failed to shift knowledge, beliefs, or peer support suggests that the binding constraint was not awareness or demand, but teachers' ability to translate DL into classroom routines under real-world conditions.

These findings have important implications for the economics of scaling. First, they demonstrate that cost-saving interventions may fail to achieve impact when they omit complementary inputs necessary for production—here, the hands-on, interactive elements of skill formation. Second, they suggest limits to digital training as a standalone tool in weak-incentive bureaucracies: without high engagement and mechanisms that approximate peer interaction and guided practice, digital platforms are unlikely to sustain complex reforms. Third, the findings generalize to many government programs where front-line workers must implement high-skill tasks but operate under low-powered incentives and limited managerial oversight.

More broadly, the paper provides an empirical framework for predicting when "scaling down" delivery models will succeed or fail. Low-cost approaches are likely to work only when the underlying task is simple, does not require tacit skill, or when digital systems are designed to recreate the core experiential ingredients that drive behavior change. In settings where tasks require specialized

human capital, high-touch support may remain essential—even when budgets are tight.

Designing scalable public-sector reforms therefore requires more than identifying cheap substitutes; it requires identifying which inputs are fundamental to execution and which can be economized without eroding impact. Understanding and addressing these constraints is central to achieving sustained improvements in public service delivery.

References

Aker, J. C., I. Ghosh, and J. Burrell (2016). The promise (and pitfalls) of ict for agriculture initiatives. *Agricultural Economics 47*(S1), 35–48.

Anderson, M. 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), 1481–1495.

Angrist, N. and R. Meager (2023). Implementation matters: Generalizing treatment effects in education. Technical report, Working Paper.

Banerjee, A., R. Banerji, J. Berry, E. Duflo, H. Kannan, S. Mukerji, M. Shotland, and M. Walton (2017). From proof of concept to scalable policies: challenges and solutions, with an application. *Journal of Economic Perspectives 31*(4), 73–102.

Banerjee, A., S. Cole, E. Duflo, and L. Linden (2007). Remedying education: Evidence from two randomized experiments in india. *The Quarterly Journal of Economics*.

Baseler, T. (2023, October). Hidden income and the perceived returns to migration. *American Economic Journal: Applied Economics 15*(4), 321–52.

Beg, S. A., A. E. Fitzpatrick, and A. Lucas (2023). Managing to learn. Technical report, National Bureau of Economic Research.

Beg, S. A., A. M. Lucas, W. Halim, and U. Saif (2019, March). Beyond the basics: Improving post-primary content delivery through classroom technology. Working Paper 25704, National Bureau of Economic Research.

Ben-Porath, Y. (1967). The production of human capital and the life cycle of earnings. *Journal of political economy 75*(4, Part 1), 352–365.

Bold, T., M. Kimenyi, G. Mwabu, A. Ng'ang'a, and J. Sandefur (2018). Experimental evidence on scaling up education reforms in kenya. *Journal of Public Economics 168*, 1–20.

Dampson, D. G. (2021). Effectiveness of professional learning communities in ghanaian basic schools through the lenses of socio-cultural theory. *Journal of Educational Issues 7*(2), 338–354.

Davies, E., P. Deffebach, L. Iacovone, and D. J. McKenzie (2024). Training microentrepreneurs over zoom: Experimental evidence from mexico. *Journal of Development Economics 167*, 103244.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher 38*(3), 181–199.

Duflo, A., J. Kiessel, and A. M. Lucas (2024, 02). Experimental evidence on four policies to increase learning at scale. *The Economic Journal 134*(661), 1985–2008.

Duflo, E., P. Dupas, and M. Kremera (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya. *The American Economic Review 101*(5), 1739–1774.

Dupas, P. (2011, January). Do teenagers respond to hiv risk information? evidence from a field experiment in kenya. *American Economic Journal: Applied Economics 3*(1), 1–34.

Escueta, M., V. Quan, A. J. Nickow, and P. Oreopoulos (2017). Education technology: An evidence-based review. National Bureau of Economic Research Working Paper 23744.

Fabregas, R., M. Kremer, and F. Schilbach (2019). Realizing the potential of digital development: The case of agricultural advice. *Science 366*(6471), eaay3038.

Free, C., G. Phillips, L. Watson, L. Galli, L. Felix, P. Edwards, V. Patel, and A. Haines (2013). The effectiveness of mobile-health technologies to improve health care service delivery processes: a systematic review and meta-analysis. *PLoS medicine 10*(1), e1001363.

Fujiwara, T. (2015). Voting technology, political responsiveness, and infant health: Evidence from brazil. *Econometrica 83*(2), 423–464.

Hanushek, E. A. and S. G. Rivkin (2010). Generalizations about using value-added measures of teacher quality. *American economic review 100*(2), 267–271.

Hennessy, S., S. D'Angelo, N. McIntyre, M. Brugha, and A. Zubairi (2022). Technology use for teacher professional development in low- and middle-income countries: A systematic review. *Computers and Education Open 3*, 100080.

Hirji, S., B. Park, E. Tsinigo, S. Beg, A. Fitzpatrick, and A. Lucas (2023). Facilitating real-time cost collection and evaluating cost-effectiveness in a multi-armed study with government partners in ghana. *Journal of Development Effectiveness 15*(1), 31–42.

Ibeneme, S., H. Karamagi, D. Muneene, K. Goswami, N. Chisaka, and J. Okeibunor (2022). Strengthening health systems using innovative digital health technologies in africa. *Frontiers in Digital Health 4*, 854339.

Jensen, R. (2010, 05). The (perceived) returns to education and the demand for schooling. *The Quarterly Journal of Economics 125*(2), 515–548.

Lazear, E. P. (2009). Firm-specific human capital: A skill-weights approach. *Journal of Political Economy 117*(5), 914–940.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies 76*(3), 1071–1102.

List, J. A. (2022). *The Voltage Effect: How to Make Good Ideas Great and Great Ideas Scale*. New York, NY: Currency.

Mehmood, M. Z. (2023). Short messages fall short for micro-entrepreneurs: Experimental evidence from kenya. Technical report, Job Market Paper, University of California, Berkeley. Version from October 17 2025; randomized field experiment with 4,700 micro-entrepreneurs in Kenya.

Muralidharan, K. and P. Niehaus (2017). Experimentation at scale. *Journal of Economic Perspectives 31*(4), 103–24.

Popova, A., D. Evans, and V. Arancibia (2022). Professional development for teachers: Evidence and policy. *World Bank Policy Research Working Paper*.

Singh, A., L. Navarro-Sola, and P. Oreopoulos (2025, December). Education technology. *VoxDevLit 20*(1).

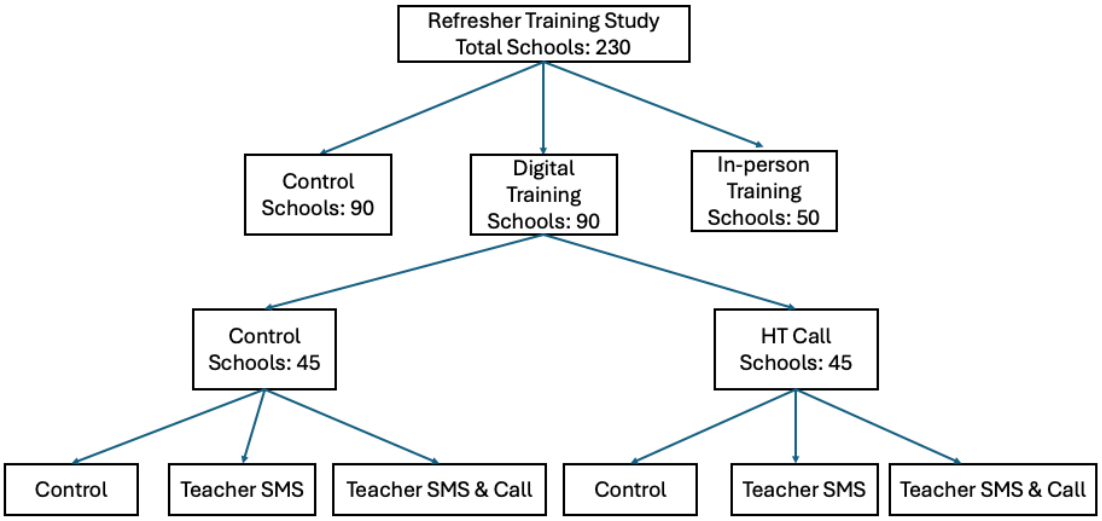
Topel, R. H. (1991). Specific capital, mobility, and wages: Wages rise with job seniority. *Journal of Political Economy 99*(1), 145–176.

Vivalt, E. (2020). How much can we generalize from impact evaluations? *Journal of the European Economic Association 18*(6), 3045–3089.

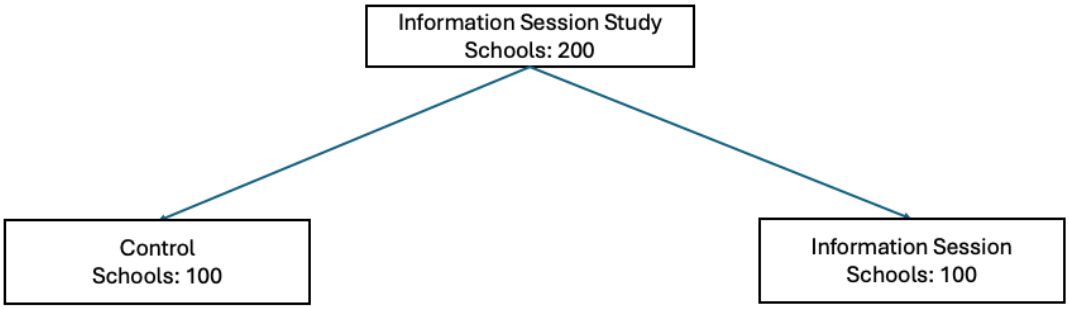
World Bank (2018). *World Development Report 2018: Learning to Realize Education's Promise*. Washington, DC: World Bank.

World Bank (2019). Ghana accountability for learning outcomes project (galop) – project appraisal document (pad3320). World Bank Documents. Project ID P165557; approved October 2, 2019.

10 Main Figures and Tables



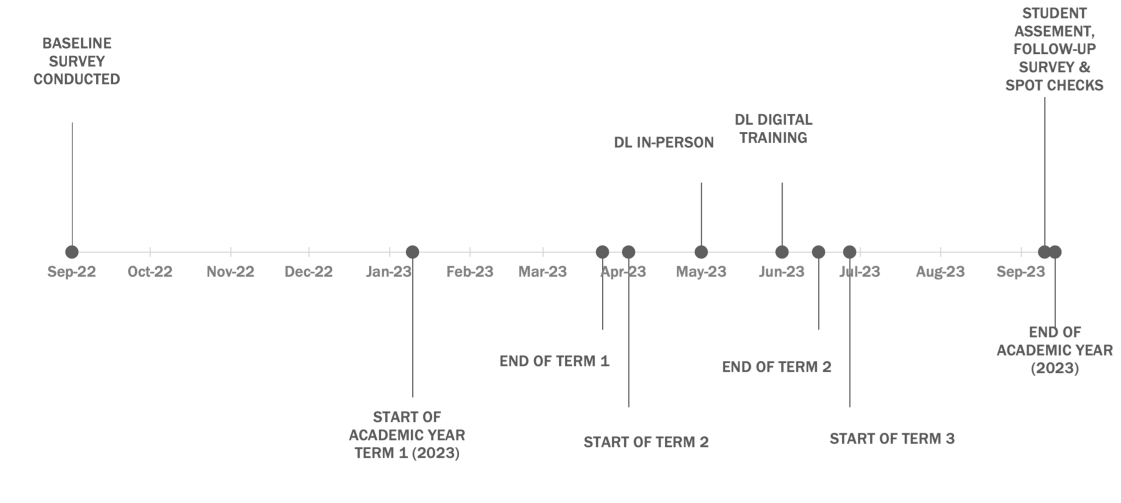
Panel A: Refresher Training Randomization



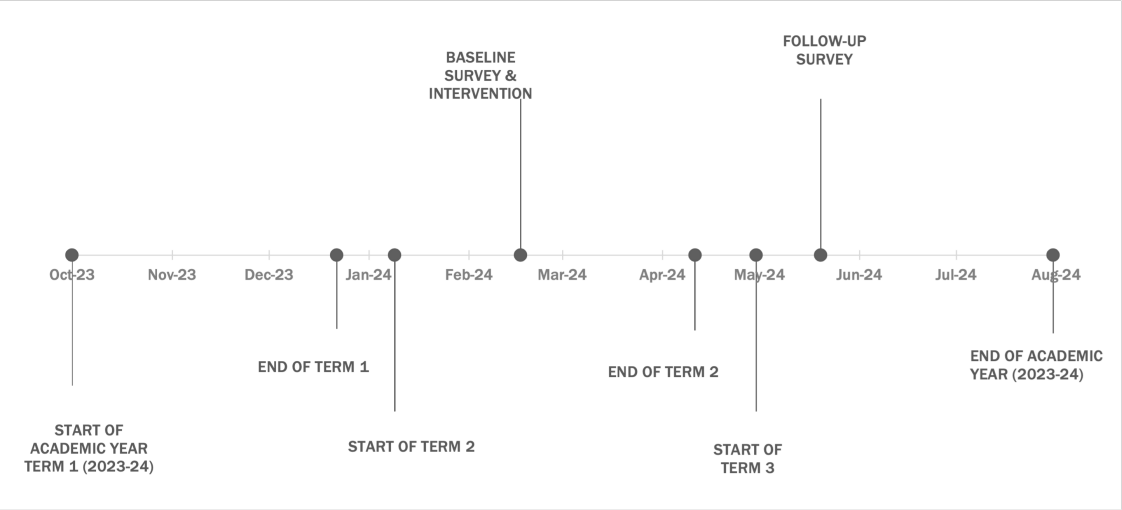
Panel B: Information Randomization

The figure shows the intended sample sizes of both studies. In the refresher training (Panel A), 5 schools were dropped between baseline and follow-up due to either closures or security issues. In addition, the sample size varies by respondent type due to vacancies and absenteeism. Within the digital training group, the intended sample was 90 schools. In the sub-randomization, there are 28 schools that could not be reached due to lack of complete contact information. Randomization is stratified by region in Panel A (both the refresher training and the sub-randomization), as well as in Panel B.

Figure 1: Randomization

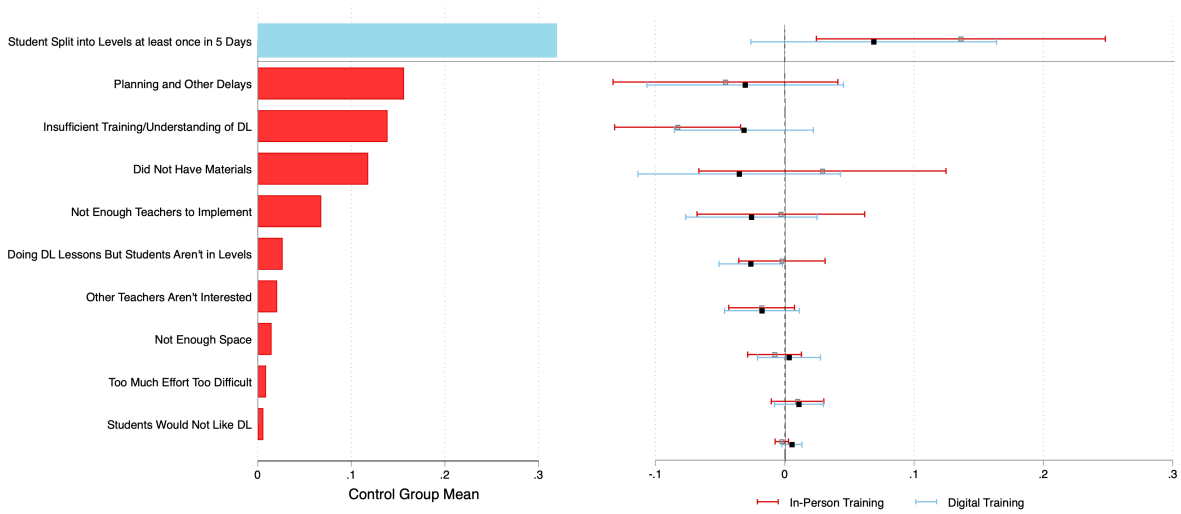


Panel A: Refresher Training Study



Panel B: Information Session Study

Figure 2: Timeline



The blue bar denotes teachers successfully implemented DL. Red bars correspond to barriers that prevented DL implementation.

Figure 3: Teacher Reported Implementation & Barriers to DL Implementation

	Knowledge of Task	Knowledge of Benefits	Signal of Priority	Skills to Implement Task	Peer Support / Community
In-Person Refresher Training	✓✓	✓✓	✓✓	✓✓	✓✓
Digital Refresher Training	✓	✓	✓	✓	✓
Information	–	✓	✓	–	–
SMS Nudges	–	–	✓	–	–
SMS Text	–	–	✓✓	–	–

Figure 4: Potential Effects of the Refresher Training, Information and Nudge Interventions

Table 1: Implementation

	Teachers' Report				Enumerator's Report
	Students Participated in Levelling Exams	**Students Split into Levels at least once in 5 Days**	**No of Days Students Split Into Levels**	**Implemented DL Today in Class**	**Class in Levels**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**
Panel A: Refresher Training					
In-Person Training	0.233***	0.136**	0.586***	0.089	0.071*
	(0.070)	(0.057)	(0.200)	(0.060)	(0.042)
Digital Training	0.098*	0.069	0.241	0.077*	0.034
	(0.056)	(0.048)	(0.153)	(0.045)	(0.027)
Observations	602	602	602	602	691
P-value: Digital = In-Person	0.077	0.277	0.102	0.848	0.392
R^2	0.101	0.176	0.183	0.080	0.047
Control Group Mean	0.280	0.319	0.735	0.180	0.028
Panel B: Information Session					
Information	-0.015	0.030	0.170	0.004	-0.023
	(0.041)	(0.040)	(0.141)	(0.043)	(0.035)
Observations	810	810	810	810	810
R^2	0.210	0.384	0.415	0.289	0.227
Control Group Mean	0.182	0.276	0.860	0.218	0.111
Panel C: SMS Text					
Peer Info SMS	-0.011	0.040	0.238	-0.010	0.017
	(0.047)	(0.052)	(0.173)	(0.052)	(0.043)
Information	-0.017	-0.069	-0.215	0.031	-0.020
	(0.047)	(0.052)	(0.173)	(0.050)	(0.043)
Observations	741	741	741	741	741
Sum of Coefficients	-0.028	-0.028	0.023	0.021	-0.003
P-value	0.701	0.699	0.924	0.763	0.951
R^2	0.195	0.277	0.268	0.215	0.115
Control Group Mean	0.293	0.388	1.108	0.269	0.108

Note: Column 1 - 5 use the sample includes teachers at endline survey while column 6 uses the enumerator's observations at the endline. Panel A reports estimate for refresher training. Panel B reports estimate for information. Covariates include age & experience (linear and square), level of class, gender and baseline implementation of DLE. The standard errors are clustered at school level. $*p < 0.10, **p < 0.05, ***p < 0.01$.

Table 2: Test Scores

	Combined Score (1)	English Score (2)	Math Score (3)
In-Person Training	0.071**	0.085**	0.048
	(0.033)	(0.035)	(0.034)
Digital Training	-0.001	-0.008	0.008
	(0.029)	(0.031)	(0.032)
P-value: Digital = In-Person	0.03	0.01	0.22
Observations	3,211	3,211	3,211
R^2	0.76	0.73	0.68
Control Group Mean	0.48	0.44	0.48

Note: Table includes the sample for students at endline. Covariates include standardized baseline scores of mathematics & english, class, gender and age of the student. The standard errors are clustered at school level. $*p < 0.10, **p < 0.05, ***p < 0.01$.

Table 3: Mechanisms: Knowledge, Peer Support and Perceptions

	Proportion of Questions Answered Correctly (1)	Ministry cares if Teacher Implements DL (2)	Received Instruction on DL from HT (3)	Received Instruction on DL from another teacher (4)	1:1 Online Session for Teaching (5)	Preferred DL over Curriculum (6)	DL should be New Standard (7)
Panel A: Refresher Training							
In-Person Training	0.020	0.003	0.026	0.026	0.069	0.107***	
	(0.015)	(0.031)	(0.051)	(0.061)	(0.044)	(0.040)	
Digital Training	0.005	0.015	0.061	0.085*	0.150***	0.021	
	(0.014)	(0.029)	(0.048)	(0.048)	(0.037)	(0.031)	
Observations	602	602	602	602	602	602	
P-value: Digital = In-Person	0.347	0.740	0.518	0.324	0.086	0.041	
R^2	0.034	0.026	0.086	0.046	0.063	0.075	
Control Group Mean	0.552	0.853	0.622	0.543	0.106	0.094	
Panel B: Information Session							
Information	-0.009	0.020	-0.014	0.081**		0.007	0.060**
	(0.012)	(0.032)	(0.044)	(0.039)		(0.031)	(0.028)
Observations	810	810	810	810		810	810
R^2	0.088	0.055	0.167	0.199		0.039	0.075
Control Group Mean	0.551	0.753	0.540	0.412		0.174	0.751
Panel C: SMS Text							
Peer Info SMS	-0.004	0.000	-0.051	-0.019		0.031	0.032
	(0.012)	(0.027)	(0.047)	(0.039)		(0.030)	(0.027)
Information	-0.003	0.031	0.045	0.029		-0.016	0.040
	(0.012)	(0.027)	(0.047)	(0.039)		(0.029)	(0.027)
Observations	741	741	741	741		741	741
Sum of Coefficients	-0.007	0.031	-0.006	0.010		0.015	0.072
P-value	0.694	0.454	0.924	0.856		0.726	0.061
R^2	0.119	0.062	0.088	0.134		0.053	0.082
Control Group Mean	0.514	0.847	0.633	0.538		0.214	0.826

Note: The sample includes teachers at endline survey. Panel A reports estimate for refresher training. Panel B reports estimate for information. Covariates include age & experience (linear and square), level of class, gender and baseline implementation of DLE. The standard errors are clustered at school level. $*p < 0.10, **p < 0.05, ***p < 0.01$.

Table 4: Mechanisms: Teachers' Ability to Teach DL

	Teacher Could Teach DL If Willing (1)	Confidence in DL Curriculum Knowledge (2)	Teachers Find Teaching DL Easy (3)	Teachers Find Conducting DL Exams Easy (4)	Teachers Use DL Materials Outside Class (5)
Panel A: Refresher Training					
In-Person Training	0.068**		0.065	0.083***	0.117**
	(0.029)		(0.052)	(0.032)	(0.058)
Digital Training	-0.012		-0.013	0.013	0.027
	(0.030)		(0.040)	(0.018)	(0.045)
Observations	602		602	602	602
P-value: Digital = In-Person	0.010		0.150	0.034	0.130
R^2	0.045		0.075	0.054	0.133
Control Group Mean	0.894		0.183	0.032	0.251
Panel B: Information Session					
Information	0.005	-0.072*	0.029	-0.034	0.042
	(0.028)	(0.038)	(0.027)	(0.021)	(0.040)
Observations	810	810	810	810	810
R^2	0.064	0.142	0.070	0.079	0.159
Control Group Mean	0.823	0.656	0.143	0.058	0.368
Panel C: SMS Text					
Peer Info SMS	0.030	0.036	0.025	0.013	0.006
	(0.025)	(0.035)	(0.034)	(0.018)	(0.042)
Information	-0.018	-0.034	0.037	0.010	0.021
	(0.024)	(0.035)	(0.033)	(0.018)	(0.044)
Observations	741	741	741	741	741
Sum of Coefficients	0.012	0.003	0.062	0.023	0.027
P-value	0.759	0.963	0.206	0.419	0.686
R^2	0.060	0.129	0.081	0.060	0.044
Control Group Mean	0.884	0.744	0.264	0.042	0.528

Note: The sample includes teachers at endline survey. Panel A reports estimate for refresher training. Panel B reports estimate for information. Covariates include age & experience (linear and square), level of class, gender and baseline implementation of DLE. The standard errors are clustered at school level. $*p < 0.10, **p < 0.05, ***p < 0.01$.

Appendix for Online Publication

A Additional Intervention Details

Digital Training. The NTC hosted the content on the website ntc.gov.gh/ipa-and-ntc/ and created user accounts for all teachers in Digital Training schools with their email addresses as their usernames. Teachers were invited to a webinar orientation session in June 2023 that focused on addressing potential challenges with platform access. The orientation walked teachers through completing their platform registration. When teachers accessed the portal for the first time, they were prompted to generate passwords, which they used for subsequent logons.

Before deploying training modules, the NTC conducted pilot testing to address potential issues, including unstable internet connectivity in remote areas. To mitigate this, teachers could download the modules for offline access.

The content was divided into 5 modules. Teachers had to pass assessment tests at the end of each module to progress to the next module. To complete the course, teachers were supposed to video themselves conducting a DL lesson and upload it to the platform. The uploads were supposed to be analyzed and scored by master trainers. Teachers did not upload these videos. To incentivize completion, teachers earned 2 CPD points upon finishing the digital training. Teachers were initially told to complete the modules within two weeks. Near the end of the two weeks, all teachers were given a one week extension. The digital portal was accessible from June 5 to June 24, 2023.

A version of the DLE digital training platform was opened up to all teachers in Ghana in the fourth quarter of 2023, reaching over 90,000 teachers. This mass engagement with the platform occurred after the conclusion of our follow-up survey.

Pamphlet and Session. The intervention was implemented through a series of group information sessions, facilitated by field staff from Innovations for Poverty Action (IPA). These sessions were scheduled to follow interviews and classroom observations in intervention schools. The session began with an introduction that emphasized the purpose of the meeting and the significance of DL within the broader context of GALOP. Teachers and HTs were then guided through the information sheet, with particular attention given to the infographics and success stories from Ghanaian teachers who have successfully implemented DL. These testimonials were included to provide real-world examples and inspire confidence in the intervention. The field staff facilitated discussions and answered any questions raised by the participants, ensuring that they fully understood the material and felt equipped to implement DL strategies effectively. The objective was to create a collaborative environment where teachers could deepen their understanding of DL and engage with the information sheets in a structured manner, having a shared understanding of DL's potential benefits and practical application in the classroom.

B Additional Tables

Tables in this section contain additional supporting materials.

Table A1: Balance Table (Refresher Training Intervention)

Variables	Control		Digital Training		In-person Training		
	Mean (1)	SD (2)	Mean (3)	SD (4)	Mean (5)	SD (6)	P-value (7)
Panel A: Teacher-Level Variables							
Female	0.31	0.46	0.34	0.48	0.30	0.46	0.65
Age	33.69	6.52	34.80	6.63	34.10	7.37	0.21
Teacher has BA	0.28	0.45	0.37	0.48	0.30	0.46	0.12
Years of experience	6.71	5.37	7.59	5.79	7.24	6.69	0.22
Students split into levels	0.57	0.50	0.53	0.50	0.44	0.50	0.25
N	435		438		243		
Panel B: Head-Teacher Level Variables							
Female	0.16	0.37	0.14	0.35	0.10	0.31	0.55
Age	41.43	8.46	43.02	7.93	39.92	7.64	0.08
Years of experience	5.58	3.97	5.85	4.71	6.56	6.31	0.50
Students split into levels	0.63	0.49	0.60	0.49	0.65	0.48	0.86
Students benefitted from DL	0.95	0.20	0.95	0.20	0.90	0.29	0.39
Teachers successfully implemented DL	0.87	0.32	0.79	0.39	0.82	0.35	0.59
N	88		88		49		
Panel C: Student Level Variables							
Age	10.97	1.93	10.77	1.95	10.81	1.95	0.48
Female	0.48	0.50	0.51	0.50	0.49	0.50	0.36
P4-P6 students taught in levels	0.60	0.49	0.5	0.49	0.60	0.49	0.95
Implementation Index	0.02	1.04	0.00	1.03	-0.01	0.87	0.84
Standardized Math Scores	0.07	0.98	-0.07	1.02	0.04	0.97	0.62
Standardized English Scores	0.08	0.96	-0.08	1.04	0.02	0.96	0.49
Class 2	0.25	0.43	0.25	0.44	0.25	0.44	0.42
Class 3	0.25	0.44	0.25	0.43	0.25	0.44	0.99
Class 4	0.25	0.43	0.25	0.44	0.25	0.43	0.74
Class 5	0.24	0.43	0.24	0.43	0.24	0.43	0.64
N	1374		1355		751		
Panel D: School Level Variables							
School has electricity	0.42	0.50	0.44	0.50	0.50	0.51	0.59
School has stable phone network	0.50	0.50	0.51	0.50	0.60	0.49	0.45
Total Students	188.3	139.8	170.1	116.9	168.5	104.3	0.55
Number of SISO visits to the school	7.06	3.55	6.64	3.39	8.04	4.04	0.15
N	88		88		49		

Note: The table reports baseline balance statistics for refresher training intervention. Panel A shows statistics for teachers. Panel B presents statistics for head teachers. Panel C reports student-level variables. Math Scores and English Scores are standardized using the entire baseline sample. Panel D shows the school level variables. Implementation Index is a standardized weighted index (Stata's swindex command) constructed from teacher responses, based on whether students were split into levels and whether students participated in the levelling exams. Columns (1), (3), and (5) report the means for the Control, Digital Training, and In-person Training groups, respectively. Columns (2), (4), and (6) report the standard deviations. Column (7) reports the p-value from a test of equality of means across the three groups. N represents the number of observations in each group for the respective panel.

Table A2: Balance Table (Information Session)

Variables	Control		Information		
	Mean (1)	SD (2)	Mean (3)	SD (4)	P-value (5)
Panel A: Teacher-Level Variables					
Female	0.39	0.49	0.41	0.49	0.78
Age	36.07	7.35	35.77	7.05	0.85
Teacher has BA	0.44	0.50	0.45	0.50	0.98
Years of experience	10.15	7.43	9.62	7.08	0.40
Students split into levels	0.24	0.43	0.25	0.43	0.93
Proportions of questions answered correctly	0.54	0.14	0.53	0.16	0.34
N	437		441		
Panel B: Head-Teacher Level Variables					
Female	0.24	0.43	0.30	0.46	0.34
Age	44.15	7.30	43.37	7.28	0.37
Years of experience	6.47	4.72	6.95	4.68	0.58
Students split into levels	0.34	0.48	0.39	0.49	0.55
Proportions of questions answered correctly	0.56	0.14	0.53	0.15	0.17
DL improves student learning	0.96	0.20	0.97	0.17	0.76
Student would enjoy DL	0.98	0.14	0.97	0.17	0.46
Takes less time for teachers to prepare DL lessons	0.39	0.49	0.35	0.48	0.58
DL is more effective than the curriculum	0.66	0.48	0.63	0.48	0.66
N	96		98		
Panel C: School Level Variables					
Student Split into Levels (Enumerator Observation)	0.08	0.24	0.08	0.24	0.99
Total Students	214.46	142.89	217.86	162.70	0.93
Number of SISO visits to the school	4.17	2.66	3.78	2.13	0.26
N	96		98		

Note: The table reports baseline balance statistics for information. Panel A shows statistics for teachers. Panel B presents statistics for head teachers while Panel C shows the school level variables. Columns (1) and (3) report the means for the control and information sheet, respectively. Columns (2) and (4) report the standard deviations. Column (5) reports the p-value from a test of equality of means between treatment & control groups. *N* represents the number of observations in each group for the respective panel.

Table A3: Balance Table (SMS Text)

Variables	Control		Peer Info SMS		
	Mean (1)	SD (2)	Mean (3)	SD (4)	P-value (5)
Panel A: Teacher-Level Variables					
Female	0.41	0.49	0.39	0.49	0.75
Age	35.57	6.91	36.28	7.47	0.22
Teacher has BA	0.44	0.50	0.46	0.50	0.71
Years of experience	9.69	7.00	10.08	7.51	0.63
Students split into levels	0.26	0.44	0.23	0.42	0.51
Proportions of questions answered correctly	0.54	0.15	0.53	0.15	0.54
N	447		432		
Panel B: Head-Teacher Level Variables					
Female	0.28	0.45	0.26	0.44	0.70
Age	43.98	7.42	43.52	7.17	0.57
Years of experience	7.13	4.68	6.28	4.70	0.15
Students split into levels	0.39	0.49	0.34	0.48	0.42
Proportions of questions answered correctly	0.56	0.14	0.53	0.15	0.13
DL improves student learning	0.98	0.14	0.95	0.22	0.16
Student would enjoy DL	0.98	0.14	0.97	0.17	0.50
Takes less time for teachers to prepare DL lessons	0.38	0.49	0.35	0.48	0.76
DL is more effective than the curriculum	0.65	0.48	0.64	0.48	0.62
N	100		100		
Panel C: School Level Variables					
Student Split into Levels (Enumerator Observation)	0.08	0.25	0.07	0.23	0.81
Total Students	217.72	128.05	214.59	175.23	0.93
Number of SISO visits to the school	3.91	2.46	4.03	2.36	0.72
N	100		100		

Note: The table reports baseline balance statistics for information. Panel A shows statistics for teachers. Panel B presents statistics for head teachers while Panel C shows the school level variables. Columns (1) and (3) report the means for the control and information sheet, respectively. Columns (2) and (4) report the standard deviations. Column (5) reports the p-value from a test of equality of means between treatment & control groups. N represents the number of observations in each group for the respective panel.

Table A4: Takeup

	Attended Training on DL at Central Location (1)	**Signed Up on Digital Training Portal (2)**
In-Person Training	0.535***	-0.027
	(0.049)	(0.056)
Digital Training	0.030	0.105**
	(0.048)	(0.048)
Observations	602	602
P-value: Digital = In-Person	0.000	0.025
R^2	0.230	0.119
Control Group Mean	0.322	0.481

Note: The sample includes teachers at endline survey. Covariates include age & experience (linear and square), level of class, gender and baseline implementation of DLE. The standard errors are clustered at school level. $*p < 0.10, **p < 0.05, ***p < 0.01$.

Table A5: Effects of Encouragement on Digital Training Engagement

	Join WhatsApp (1)	Completed Pre-Course Survey (2)	Completed Post-Course Survey (3)	Course Total Real (4)
No HT call but teacher received SMS (2)	0.023	-0.014	0.071	4.492
	(0.068)	(0.057)	(0.052)	(4.498)
No HT call but teacher received SMS & call (3)	0.061	0.008	0.092*	4.018
	(0.060)	(0.047)	(0.053)	(4.587)
Got HT call but teacher neither received SMS nor call (4)	-0.050	-0.105*	0.017	-3.717
	(0.056)	(0.055)	(0.033)	(3.232)
Got HT call and teacher received SMS (5)	-0.013	-0.049	0.036	-3.074
	(0.059)	(0.061)	(0.037)	(3.088)
Got HT call and teacher received SMS & call (6)	0.002	0.001	0.033	3.447
	(0.059)	(0.066)	(0.029)	(5.250)
2 vs 3	0.57	0.69	0.61	0.90
2 vs 4	0.36	0.24	0.37	0.20
2 vs 5	0.66	0.67	0.57	0.23
2 vs 6	0.80	0.87	0.51	0.89
3 vs 4	0.11	0.05	0.19	0.18
3 vs 5	0.30	0.37	0.34	0.20
3 vs 6	0.41	0.91	0.28	0.94
4 vs 5	0.53	0.13	0.53	0.78
4 vs 6	0.09	0.06	0.60	0.07
5 vs 6	0.82	0.45	0.95	0.19
Observations	340	340	340	340
R^2	0.07	0.07	0.04	0.08
Control Group Mean	0.10	0.12	0.02	5.13
Overall Mean	0.11	0.09	0.06	6.12
Overall SD	0.31	0.29	0.24	22.73

Note: The estimates are created using a sample of data which consists of subset of teachers that were invited to complete digital training modules. Course Total Real is graded out of 120 scores. All regressions include strata (region) fixed effects, and gender of teacher as controls. The standard errors are clustered at school levels. $*p < 0.10, **p < 0.05, ***p < 0.01$

Table A6: Belief Assessment in the Information Session Study

	DL Improves Student Learning (1)	DL is More Effective than Traditional Learning (2)	Percentage of Students Who Can Read 3 Letter Words (3)
Panel A: Immediately After Intervention			
Post - Information Sheet	0.016	0.098**	0.180***
	(0.012)	(0.041)	(0.026)
Observations	386	386	386
R^2	0.05	0.09	0.30
Pre Outcome Mean	0.98	0.70	0.54
Panel B: Endline			
Post - Information Sheet	0.013	0.026	0.035
	(0.024)	(0.044)	(0.039)
Observations	1019	1019	1019
R^2	0.60	0.72	0.70
Control Group Mean	0.95	0.67	0.54

Note: Panel A reports the estimates from the sample that includes the schools matched at class level between the baseline teacher & head teacher survey and amenities survey. All columns use class levels and region fix effects as covariates. For Panel B, the sample is prepared using the endline and baseline teacher & HT survey. All columns use class levels, teacher fix effects and region fix effects as covariates. The standard errors are clustered at school level. $*p < 0.10, **p < 0.05, ***p < 0.01$.

Table A7: Implementation of DLE (Reported by Head Teacher)

	DL Implemented in P2-P3 (1)	DL Implemented in P4-P6 (2)
Panel A: Refresher Training		
In-Person Training	0.070	0.151*
	(0.090)	(0.086)
Digital Training	-0.029	0.011
	(0.079)	(0.077)
Observations	218	218
P-value: Digital = In-Person	0.281	0.120
R^2	0.084	0.124
Control Group Mean	0.460	0.437
Panel B: Information Session		
Information	0.002	-0.022
	(0.048)	(0.047)
Observations	188	188
R^2	0.627	0.637
Control Group Mean	0.391	0.402
Panel C: SMS Text		
Peer Info SMS	0.040	0.009
	(0.064)	(0.063)
Information	-0.005	-0.039
	(0.065)	(0.064)
Observations	185	185
Sum of Coefficients	0.035	-0.030
P-value	0.688	0.716
R^2	0.345	0.373
Control Group Mean	0.500	0.511

Note: The sample includes head teachers at endline survey. Panel A reports estimate for refresher training. Panel B reports estimate for information. Covariates include age & experience (linear and square), gender and baseline implementation of DLE. The standard errors are clustered at school level. $*p < 0.10, **p < 0.05, ***p < 0.01$

Table A8: Implementation (Multiple Hypothesis Testing)

	Teachers' Report				Enumerator's Report
	Students Participated in Levelling Exams (1)	**Students Split Into Levels For Atleast Once in 5 Days (2)**	**No of Days Students Split Into Levels (3)**	**Teach DL Today (4)**	**Students Levelled (5)**
Panel A: Refresher Training Intervention					
In-Person Training	0.233	0.136	0.586	0.089	0.071
	(0.001)	(0.017)	(0.004)	(0.139)	(0.092)
	[0.023]	[0.139]	[0.046]	[0.402]	[0.372]
Digital Training	0.098	0.069	0.241	0.077	0.034
	(0.080)	(0.155)	(0.118)	(0.091)	(0.221)
	[0.372]	[0.416]	[0.381]	[0.372]	[0.504]
Panel B: Information Session					
Information	-0.015	0.030	0.170	0.004	-0.023
	(0.722)	(0.465)	(0.229)	(0.934)	(0.516)
	[0.987]	[0.987]	[0.901]	[0.990]	[0.987]

Note: Column 1 - 4 use the sample includes teachers at endline survey while column 5 uses the enumerator's observations at the endline. P-values are reported in parenthesis while q-values are reported in square brackets. Panel A reports estimate for refresher training. Panel B reports estimate for information. Covariates include age & experience (linear and square), level of class, gender and baseline implementation of DLE. The standard errors are clustered at school level.

Table A9: Attendance & Test Scores (Lee Bounds)

	Present The Day (1)	Took Test (2)	Combined Score Lower Bound (3)	Combined Score Upper Bound (4)	English Score Lower Bound (5)	English Score Upper Bound (6)	Math Score Lower Bound (7)	Math Score Upper Bound (8)
In-Person Training	-0.024	-0.011	0.060*	0.059*	0.075**	0.075**	0.038	0.038
	(0.033)	(0.013)	(0.033)	(0.031)	(0.035)	(0.033)	(0.034)	(0.033)
Digital Training	-0.041	-0.029***	-0.010	-0.006	-0.017	-0.010	-0.001	0.004
	(0.027)	(0.011)	(0.030)	(0.029)	(0.031)	(0.031)	(0.032)	(0.031)
Observations	3,450	3,450	3,140	3,140	3,140	3,140	3,140	3,140
R^2	0.05	0.02	0.75	0.75	0.72	0.72	0.68	0.67

Note: Lee (2009) bounding method applied to main analysis sample of test scores. Table includes the sample for students at endline. Covariates include standardized baseline scores of mathematics & english, class, gender and age of the student. The standard errors are clustered at school level. $*p < 0.10, **p < 0.05, ***p < 0.01$

Table A10: Took Test Robustness

	Present The Day (1)	Took Test (2)	Took Test (3)
In-Person Training	-0.021	-0.012	-0.012
	(0.033)	(0.014)	(0.014)
Digital Training	-0.037	-0.027**	-0.027**
	(0.027)	(0.011)	(0.011)
Standardized Math Score Baseline	0.028*	0.015*	0.019
	(0.016)	(0.009)	(0.013)
Standardized English Score Baseline	0.029*	0.004	-0.004
	(0.016)	(0.009)	(0.012)
Digital Training x Standardized Math Scores Baseline			0.002
			(0.018)
Digital Training x Standardized English Scores Baseline			0.007
			(0.018)
In-Person Training x Standardized Math Scores Baseline			-0.019
			(0.023)
In-Person Training x Standardized English Scores Baseline			0.022
			(0.024)
P-value: Digital = In-Person	0.61	0.31	0.30
Observations	3,480	3,480	3,480
R^2	0.05	0.02	0.02
Mean Dep., Control	0.64	0.94	0.94

Note: Table includes the sample for students at endline. Covariates include standardized baseline scores of mathematics & english, class, gender and age of the student. The standard errors are clustered at school level. $*p < 0.10, **p < 0.05, ***p < 0.01$.

Table A11: Teachers' Attendance

	Teacher Still Active (1)	**Present the Day (2)**
Panel A: Refresher Training		
In-Person Training	0.090**	0.054
	(0.035)	(0.066)
Digital Training	0.041	0.032
	(0.031)	(0.055)
Observations	870	660
P-value: Digital = In-Person	0.163	0.722
R^2	0.015	0.114
Control Group Mean	0.722	0.541
Panel B: Information Session		
Information	0.004	0.025
	(0.019)	(0.038)
Observations	846	787
R^2	0.075	0.133
Control Group Mean	0.931	0.687

Note: The sample includes teachers using arrival survey. Panel A reports estimate for refresher training. Panel B reports estimate for information. Covariates include age & experience (linear and square), gender and baseline implementation of DLE. The standard errors are clustered at school level. $*p < 0.10, **p < 0.05, ***p < 0.01$

Table A12: Training Usefulness & Material Access

	Teachers find in-person training useful **(1)**	**Teachers receive teaching & learning materials** **(2)**
Panel A: Refresher Training		
In-Person Training	0.193***	0.118**
	(0.040)	(0.052)
Digital Training	0.005	0.016
	(0.047)	(0.043)
Observations	602	602
P-value: Digital = In-Person	0.000	0.052
R^2	0.067	0.070
Control Group Mean	0.678	0.301
Panel B: Information Session		
Information		-0.033
		(0.032)
Observations		810
R^2		0.083
Control Group Mean		0.228
Panel C: SMS Text		
Peer Info SMS		0.065*
		(0.037)
Information		0.023
		(0.038)
Observations		741
Sum of Coefficients		0.088
P-value		0.092
R^2		0.077
Control Group Mean		0.330

Note: The sample includes teachers at endline survey. Panel A reports estimate for refresher training. Panel B reports estimate for information. Covariates include age & experience (linear and square), level of class, gender and baseline implementation of DLE. The standard errors are clustered at school level. $*p < 0.10, **p < 0.05, ***p < 0.01$.

Table A13: Heterogeneity in Digital Take-up

	Logged onto Digital Training Module (1)	Logged onto Digital Training Module (2)	Logged onto Digital Training Module (3)	Logged onto Digital Training Module (4)	Logged onto Digital Training Module (5)	Logged onto Digital Training Module (6)
In-Person Training	0.046	0.007	-0.079	-0.081	-0.136*	-0.030
	(0.075)	(0.075)	(0.090)	(0.090)	(0.081)	(0.094)
Digital Training	0.084	0.151**	0.107	0.030	0.040	0.109
	(0.059)	(0.065)	(0.070)	(0.076)	(0.067)	(0.072)
Female	-0.040	-0.071	-0.070	-0.080	-0.072	-0.070
	(0.073)	(0.049)	(0.050)	(0.049)	(0.050)	(0.050)
Above Median Experience		-0.082				
		(0.090)				
Baseline Implementation	0.035	0.038	0.032	0.045	0.020	0.036
	(0.043)	(0.043)	(0.070)	(0.044)	(0.043)	(0.045)
Above Median Amenities Index				-0.038		
				(0.070)		
School has electricity					-0.030	
					(0.071)	
School has stable phone network						0.085
						(0.071)
In-Person Training x Female	-0.203*					
	(0.111)					
Digital Training x Female	0.020					
	(0.099)					
In-Person Training x Experience (Above Median)		-0.052				
		(0.116)				
Digital Training x Experience (Above Median)		-0.129				
		(0.090)				
In-Person Training x Baseline Implementation			0.101			
			(0.120)			
Digital Training x Baseline Implementation			-0.039			
			(0.098)			
In-Person Training x Amenities Index (Above Median)				0.107		
				(0.120)		
Digital Training x Amenities Index (Above Median)				0.105		
				(0.100)		
Digital Training x School has electricity					0.095	
					(0.099)	
In-Person Training x School has electricity					0.213*	
					(0.116)	
Digital Training x School has stable phone network						-0.041
						(0.101)
In-Person Training x School has stable phone network						0.007
						(0.122)
Observations	602	602	602	602	596	596
R^2	0.101	0.103	0.097	0.098	0.103	0.099
Interacted Variable	Female	Experience (Above Median)	Baseline Implementation	Amenities Index (Above Median)	Electricity Connection	Stable Phone Network
P-value: In-Person Training + In-Person Training x Interacted Variable = 0	0.056	0.623	0.782	0.741	0.352	0.767
P-value: Digital Training + Digital Training x Interacted Variable = 0	0.211	0.755	0.319	0.037	0.063	0.319

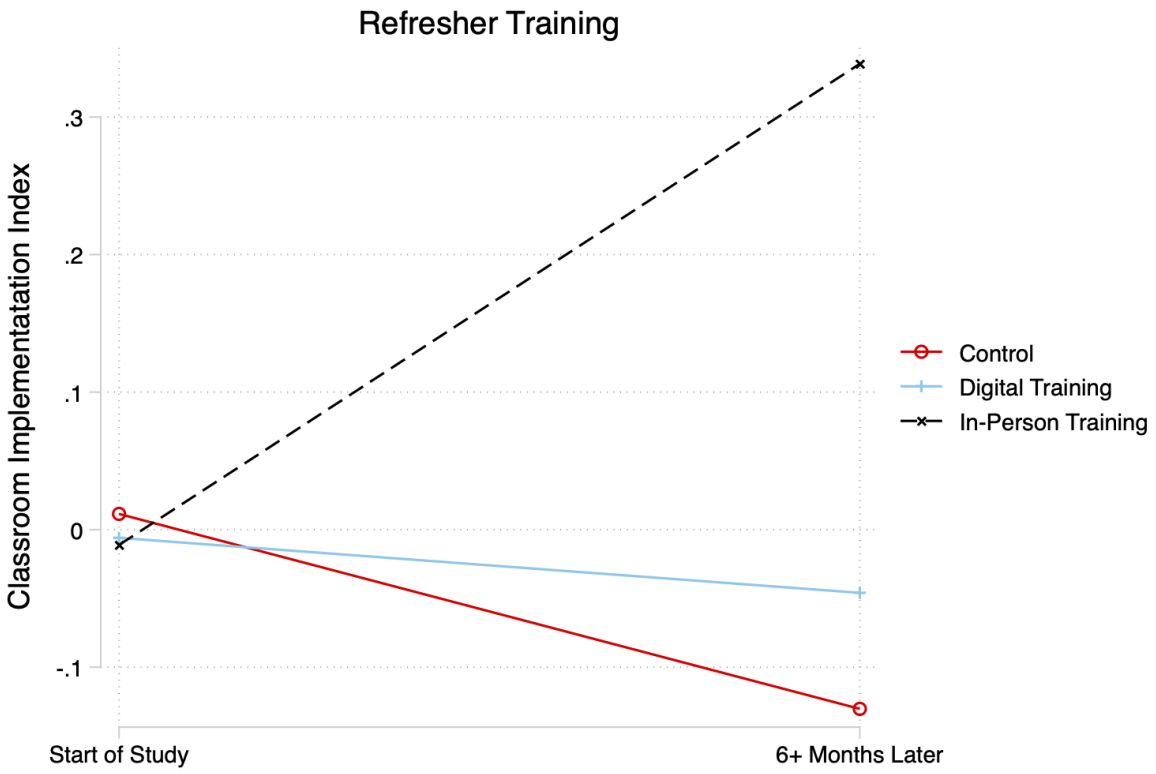
Note: The sample includes teachers at endline survey. Panel A reports estimate for refresher training. Panel B reports estimate for information. Covariates include age & experience (linear and square), level of class, gender and baseline implementation of DLE. The standard errors are clustered at school level. Rows reporting 'P-value: X + X x Interacted Variable = 0' show the p-value from a test of whether the sum of the relevant main effect and interaction coefficient equals zero. $*p < 0.10, **p < 0.05, ***p < 0.01$.

Table A14: Heterogeneity in Implementation

	Implementation (1)	Implementation (2)	Implementation (3)	Implementation (4)
Panel A: Refresher Training				
In-Person Training	0.521***	0.580***	0.189	0.611***
	(0.145)	(0.173)	(0.187)	(0.224)
Digital Training	0.226*	0.213	0.112	0.397**
	(0.125)	(0.140)	(0.142)	(0.169)
Female	-0.076	-0.098	-0.089	-0.081
	(0.159)	(0.101)	(0.102)	(0.102)
Above Median Experience		-0.022		
		(0.171)		
Baseline Implementation	0.674***	0.665***	0.460***	0.661***
	(0.102)	(0.102)	(0.158)	(0.105)
Above Median Amenities Index				0.026
				(0.163)
In-Person Training x Female	-0.176			
	(0.231)			
Digital Training x Female	0.062			
	(0.198)			
In-Person Training x Experience (Above Median)		-0.281		
		(0.211)		
Digital Training x Experience (Above Median)		0.065		
		(0.177)		
In-Person Training x Baseline Implementation			0.499*	
			(0.260)	
Digital Training x Baseline Implementation			0.290	
			(0.230)	
In-Person Training x Amenities Index (Above Median)				-0.239
				(0.280)
Digital Training x Amenities Index (Above Median)				-0.263
				(0.226)
Observations	602	602	602	602
R^2	0.163	0.165	0.170	0.168
Interacted Variable	Female	Experience (Above Median)	Baseline Implementation	Amenities Index (Above Median)
P-value: In-Person Training + In-Person Training x Interacted Variable = 0	0.111	0.064	0.000	0.025
P-value: Digital Training + Digital Training x Interacted Variable = 0	0.112	0.060	0.026	0.372
r2	0.163	0.165	0.170	0.168
Panel B: Information Session				
Information Sheet	0.070	0.013	-0.046	0.120
	(0.107)	(0.118)	(0.102)	(0.148)
Female	0.187*	0.123	0.114	0.122
	(0.111)	(0.075)	(0.076)	(0.075)
Above Median Experience		-0.005		
		(0.143)		
Baseline Implementation	1.211***	1.204***	1.097***	1.190***
	(0.144)	(0.144)	(0.202)	(0.144)
Above Median Amenities Index				0.203
				(0.147)
Information Sheet x Female	-0.132			
	(0.150)			
Information Sheet x Experience (Above Median)		0.008		
		(0.127)		
Information Sheet x Baseline Implementation			0.222	
			(0.284)	
Information Sheet x Amenities Index (Above Median)				-0.196
				(0.209)
Observations	810	810	810	810
R^2	0.342	0.341	0.343	0.346
Interacted Variable	Female	Experience (Above Median)	Baseline Implementation	Amenities Index (Above Median)
P-value: Information Sheet + Information Sheet x Interacted Variable = 0	0.690	0.867	0.507	0.604

Note: The sample includes teachers at endline survey. Panel A reports estimate for refresher training. Panel B reports estimate for information. Covariates include age & experience (linear and square), level of class, gender and baseline implementation of DLE. The standard errors are clustered at school level. $*p < 0.10, **p < 0.05, ***p < 0.01$.

C Additional Figures



Implementation Index is a standardized weighted index (Stata's swindex command) constructed from teacher responses, based on whether students were split into levels and whether students participated in the levelling exams.

Figure A1: Implementation Score by Treatment