# Making Assessment Policies Work: Bridging the Policy–Practice Gap through a Low-Cost School-Level Intervention*

Daniel Morales†
Catherine Rodríguez Orgales‡
Carlos Schmidt-Padilla§

September 26, 2026

**Abstract**

A central challenge in public policy is closing the gap between policy adoption and front-line implementation. Policies promoting the use of student-assessment information are no exception. This paper studies whether a low-cost communication technology can close that gap. We embedded a multi-arm randomized controlled trial in the rollout of a nationwide policy in the Dominican Republic requiring school principals to use census-based diagnostic assessment results to conduct workshops and develop school improvement plans. The experiment varied the number and content of phone reminders sent to principals. In the absence of reminders, implementation was limited. Only half of principals opened the official email, about a third downloaded their school reports, 11 percent documented having held at least one workshop, and fewer than 5 percent submitted an improvement plan. Assignment to a reminder treatment increased these outcomes by 19, 14, 27, and 8 percentage points, respectively. Impacts across stages of implementation varied with both the number and content of the reminders. Follow-up testing, which the COVID-19 pandemic pushed to four and five years after the intervention, provides suggestive, though less robust, evidence that these effects extended to student learning. At a direct personnel cost of about 2 dollars per school, these findings show that simple communication can substantially narrow the gap between policy mandates and school-level implementation.

*We are grateful to the Directorate for Quality Evaluation of the Dominican Republic's Ministry of Education, especially Ancell Scheker Mendoza, for its sustained support throughout this project, and to the Dominican Institute for Educational Quality Evaluation and Research for its invaluable collaboration during the study, including Julio Valeirón for project support and Ambar Tomás and the Call Center team for implementing the intervention and ensuring adherence to study protocols. We thank Uyanga Byambaa, as well as conference participants at the KDI-World Bank Development Impact Conference, for valuable feedback and suggestions. We are grateful to Noel Rodríguez for excellent research assistance. This study was registered in the American Economic Association's registry for randomized controlled trials (AEARCTR-0004954). AI-assisted tools (Claude Code and OpenAI Codex) were used to review and refine code and text originally produced by the authors. The authors independently verified all analyses, code, and written content. All errors are our own.

†Full Professor, Business School, Tecnologico de Monterrey. Email: daniel.morales@tec.mx.

‡Researcher, Department of Economics, Universidad de los Andes. Email: cathrodr@uniandes.edu.co.

§Assistant Professor, Goldman School of Public Policy, University of California, Berkeley. Email: cschmidt-padilla@gmail.com.

# 1 Introduction

A central challenge in education policy is not only identifying effective reforms, but ensuring that they are correctly implemented on the ground. Even well-designed policies or programs may have limited effects if they are not translated into concrete actions by schools, principals, and teachers. Consistent with this concern, Angrist and Dercon (2024) document large policy-practice gaps in the education sector across 50 countries and argue that ineffective service delivery, rather than poor policy design alone, is often a central bottleneck. Addressing the global learning crisis, they argue, requires developing an "implementation science" for education and treating execution, not only policy choice, as a first-order problem. Angrist et al. (2026) set out a research agenda and conceptual framework for that science. This paper studies one implementation technology, a low-cost way for a ministry to get what it mandates carried out in its schools.

The use of student-assessment information is one domain in which this implementation challenge is particularly salient. Governments around the world invest substantial resources in national and international assessments designed to measure what children learn, and the scale of these efforts has expanded rapidly over time.[1] These assessments have made clear that, despite major gains in school access across the developing world, learning levels remain strikingly low. In OECD countries they also show a sustained decline in learning, and average performance in PISA 2025 was the lowest on record in mathematics, reading and science (OECD, 2026). Yet little is known about whether schools, principals, and teachers use the information these assessments generate or whether such use can improve teaching and learning.

This paper studies these questions as a problem of implementation in the context of a national diagnostic-assessment policy in the Dominican Republic. In 2017, the Ministry of Education (MINERD) introduced the National Diagnostic Evaluations (*Evaluaciones Diagnósticas Nacionales*, EDN), a series of census-based low-stakes assessments designed to provide schools with information on student learning in the third and sixth grades. Since their introduction, MINERD has prepared school-specific reports and tasked regional and district officials with promoting their use at the school level, as required by national policy. In practice, however, it was unclear whether the availability of diagnostic reports and supporting materials was sufficient for schools to access, interpret, or use this information.

Within this setting, in partnership with MINERD and the Dominican Institute for Educational Quality Evaluation and Research (*Instituto Dominicano de Evaluación e Investigación de la Calidad Educativa*, IDEICE), we embedded a randomized controlled trial (RCT) in the 2019 national dissemination strategy of the EDN, to test whether low-cost implementation support could move schools from receiving information to acting on it. Under the national dissemination protocol, all eligible primary-school principals were sent an official email from MINERD with instructions for

[1] Cheng and Gale (2014), for example, document 306 national assessments across 85 school systems between 2004 and 2013. Just six years later, UNESCO's Inventory of Learning Assessments (UNESCO Institute for Statistics, 2025) records 1,034 assessments worldwide in 2019, nearly two-thirds of them in low- and middle-income countries.

implementing the policy. Schools were asked to download their school-specific reports through a designated link, review their results, organize three workshops—with the school leadership team, teachers, and families—and prepare an improvement plan ahead of the next assessment round in 2020. We randomly assigned principals to reminder calls encouraging them to use the reports and complete these required tasks. The strategy was designed to strengthen compliance with the requirements established by the policy, measure the extent to which they were implemented in practice, and test alternative ways of increasing implementation. The intervention varied both reminder intensity and message content, enabling us to distinguish whether reminders work primarily by directing general attention to the policy or by heightening the salience of specific implementation tasks. We combine this experimental variation with rich administrative and purpose-collected data measuring each step of the implementation chain—report access, school-level workshops, improvement plans, and the academic content of those plans—as well as subsequent student learning.

We find strikingly low levels of compliance in the absence of additional support. Data from the digital platform used to send the campaign emails show that only about half of principals in the control group opened the Ministry's email, approximately 30 percent clicked on at least one link containing materials needed to implement the policy, and a similar share downloaded their school-specific report. Reminder calls substantially increased all three measures of initial engagement, raising the probability of opening the email by 19 percentage points, clicking on a link by 25 percentage points, and downloading the school report by 14 percentage points. Effects were larger among schools assigned more calls, although marginal gains declined with call intensity for these early engagement outcomes. Message content also mattered, with calls designed to verify receipt of the information and principals' understanding of the required tasks proving especially effective at generating initial engagement.

Implementation was even lower for the more substantive actions required by the policy. Only 11 percent of control schools submitted evidence that they had conducted at least one workshop, and only 5 percent submitted a school improvement plan. Reminder calls nevertheless produced large improvements on both outcomes, increasing the probability of submitting workshop evidence by 27 percentage points and the probability of submitting an improvement plan by 8 percentage points. The intervention also increased a composite index of improvement-plan completion and quality, which averages seven dimensions of a plan's content, among them whether it used information from the diagnostic-assessment reports and whether it proposed concrete, actionable responses, and which scores a missing plan as zero. Reminder calls increased the index by 0.044, relative to a control-group mean of 0.025. Here too, reminder intensity and message content shaped these outcomes. Reminders focused on leadership-team workshops produced the largest effects on workshop-evidence submission, whereas reminders focused on teacher workshops produced the largest effects on the submission and quality of improvement plans.

Together, the findings show that implementation support through simple reminders does more than draw attention to a policy. Its design shapes whether and how far schools progress

from receiving information to translating it into concrete actions. Even so, compliance with the most demanding tasks remained low. Among schools assigned to reminder calls, only 13 percent submitted an improvement plan, and fewer than a third of those plans made substantive use of the diagnostic results.

We next ask whether these improvements in engagement and implementation translated into higher student achievement. The following rounds of national diagnostic assessments were originally scheduled for May 2020 and 2021, but the COVID-19 pandemic suspended testing until 2023 and 2024, four and five years after the intervention. We use these later assessments, administered to subsequent student cohorts, to estimate longer-run effects on achievement. Our difference-in-differences estimates indicate that the intervention increased mathematics achievement in upper-primary grades by approximately 0.05 standard deviations, with little evidence of effects in language or third grade. The estimate is smaller and imprecise under an ANCOVA specification that estimates, rather than imposes, the persistence of baseline achievement, so we read it as suggestive. The interval, the change of cohorts, and the pandemic warrant caution, but the results are consistent with the possibility that helping schools translate performance information into concrete action generates learning gains that extend to subsequent cohorts of students.

This paper contributes to several strands of the economics literature. First, it builds on a small but growing body of work examining whether providing schools with information from low-stakes standardized assessments can improve student learning (de Hoyos et al., 2017, 2021, 2024; Muralidharan and Singh, 2025), and on the broader evidence that test-score information changes behavior in education markets when it reaches parents and schools (Andrabi et al., 2017), but that information alone rarely translates into school-level action (Banerjee et al., 2010). In these studies, however, information provision was bundled with workshops, training, or other forms of implementation support, making it difficult to isolate the role of diagnostic information from the complementary inputs designed to help schools act on it. Moreover, this prior work provides limited evidence on whether schools actually used the information they received, and even less on the quality of the actions that followed. This distinction is important because weak or purely administrative implementation may help explain why some of these information-based interventions generated sizable learning gains while others did not, as discussed by Muralidharan and Singh (2025). Our evidence suggests that low-stakes diagnostic assessments can improve learning without intensive training or high-stakes accountability when schools are helped to act on the results. This also distinguishes our setting from high-stakes accountability contexts, primarily studied in developed countries, where information on test performance is often tied to sanctions, rewards, or performance-based incentives.

Second, our paper contributes to the literature on public-sector implementation, bureaucratic performance, and policy delivery at scale in developing countries reviewed by Finan et al. (2017), Pepinsky et al. (2017) and Besley et al. (2022). Much of this literature has focused on personnel selection, financial incentives, monitoring, career concerns, or formal accountability mechanisms, tools that can be powerful but are often costly, politically difficult, or administratively demand-

ing to implement at scale. We study a lighter implementation technology, direct and repeated communication from the ministry to the school principal, which requires no change in personnel, pay or formal accountability. Our paper also speaks to the literature on scaling and voltage effects, which asks why interventions that are effective in trials often lose effectiveness when incorporated into routine government systems (Bold et al., 2018; List, 2022; Duflo et al., 2024). Angrist et al. (2026) observe that studies in education are often small and tightly controlled, and that they rarely measure or report implementation fidelity and take-up. We measure take-up at every step of a national mandate, from the opening of MINERD's email to the content of the improvement plan, for the 2,647 schools in an experiment embedded in the national dissemination of the policy.

Within this broader implementation and scale-up literature, our work is closest to Agte et al. (2026), Andersen et al. (2026) and Cilliers and Habyarimana (2026), who also examine whether low-cost reminders can improve the implementation or adoption of education programs on the ground. Our design extends this evidence in four ways. First, while these studies analyze the scale-up or adoption of programs previously evaluated in randomized trials, we study an already-mandated national policy that had been in place before the intervention took place. Second, our setting includes a pure control group, which allows us to document status quo implementation levels and provide direct evidence on the implementation gap emphasized by Angrist and Dercon (2024). Third, the implementation margin we study is substantially more complex. In their settings, implementation required relatively discrete actions, such as playing a short video in class or delivering materials to families as part of a reading program. In ours, principals had to download and interpret diagnostic reports, convene leadership teams, organize workshops with teachers and families, and translate those discussions into written improvement plans. The fact that reminders remain effective in this richer setting suggests that limited attention is a binding constraint not only for one-off compliance tasks, but also for longer chains of school-level processes involving multiple actors.

Fourth, rather than studying the effect of a single reminder, we experimentally vary both the number of reminder calls and the content of the message delivered. To our knowledge, existing studies in this literature have not separately identified the marginal effects of reminder intensity and message content in a setting where implementation requires multiple sequential tasks. Although message content varied in Agte et al. (2026), the tasks required of school directors were relatively discrete, making it more difficult to distinguish whether effects were driven by the reminder itself or by the specific content of the message. In our more complex setting, we show that content matters, although the calls were timed to the task they named, so content and timing move together. Reminders that emphasize different tasks affect different points along the implementation chain, providing evidence that message framing can shape not only whether schools respond, but also how they respond.

Finally, our design is naturally related to the behavioral literature on reminders, salience, and message framing. In particular, our design speaks to a growing literature showing that limited

attention and other behavioral frictions may prevent agents from acting on information or complying with their own goals or those asked by others even when the required actions are feasible. Studies in this literature evaluating the impact of text reminders on tax compliance, savings, income reporting, immunization, and parental engagement in schooling have shown that both the timing and content of reminders can affect behavior (Antinyan et al., 2021; Karlan et al., 2016; Rodríguez and Saavedra, 2019; Banerjee et al., 2025; Bergman and Chan, 2021). Among these studies, Dustan et al. (2023) is the closest to ours, as the authors also use behavioral messages to improve the implementation of complex tasks by civil servants in the Peruvian education sector. They find that SMS messages increased compliance, but that different message contents had broadly similar effects, suggesting that timely reminders were the main driver of the response. In our setting, by contrast, both reminder intensity and message content matter. This comparison should be interpreted with some caution, however, because the differences in results may reflect several features of the intervention and setting. In particular, our reminders were delivered through phone calls rather than text messages, which may have made message content more salient, more personalized, or more difficult to ignore.

The remainder of the paper is organized as follows. Section 2 describes the institutional context and the experimental design. Section 3 presents the data and empirical strategy. Section 4 examines the effects of reminder calls on initial engagement, implementation, and learning. Section 5 concludes.

# 2 Institutional Setting and the Experimental Design

## 2.1 Learning and Assessment in the Dominican Education System

The Dominican education system is organized into two levels, primary and secondary, each comprising six grades. Centrally administered by MINERD, it is divided into 18 regions and 122 districts, and it enrolled about 2.6 million students in the 2019–2020 academic year. In 2019 the population aged 15 and older had completed 9.4 years of schooling on average (MINERD, 2021). Despite rapid economic growth and substantial poverty reduction over the last two decades, the Dominican Republic continues to underperform in education. International assessments indicate that student achievement falls short of what would be expected given the country's income level. In PISA 2025, for example, Dominican students scored 339 points in mathematics against an OECD average of 463, and only 9 percent of them reached the baseline level of proficiency, compared with 65 percent across the OECD (OECD, 2026). In response, policymakers have introduced a range of centrally designed interventions aimed at improving both access and learning.

Over the last decade, the Dominican Republic has steadily expanded and consolidated its educational assessment system. In addition to participating in international assessments, the country has institutionalized a national framework for measuring student learning built around two main instruments. The first is the Pruebas Nacionales, exit examinations administered to all students at the end of secondary school. These are high-stakes assessments, particularly for

students, since high school graduation and, in part, access to tertiary education depend on their results. The second is the EDN, census-based assessments administered in grade 3 and grade 6 of primary school and in grade 3 of secondary school. One of the three grades is assessed each year, so each is assessed once every three years. In grade 3 of primary school, these assessments cover Spanish and mathematics. In grade 6 of primary school and grade 3 of secondary school, they also cover natural science and social science. Unlike the Pruebas Nacionales, the EDN have no direct consequences for student promotion and generate no explicit incentives for schools, principals, or teachers. Instead, they are designed to provide system actors with information on student learning levels, making them a low-stakes instrument within the education system. All assessments are administered by the Dirección de Pruebas Nacionales, aligned with the national curriculum, and centrally supervised and graded to ensure comparability across schools and over time.

## 2.2 The Policy Mandate: Using Assessment Information to Improve Learning

Recognizing the potential value of assessment data, policymakers in the Dominican Republic have increasingly emphasized their use as an input into school improvement efforts. This objective is explicitly embedded in the country's legal framework. Ordinance 1-2016 requires education actors to disseminate and use assessment results as part of a broader strategy of continuous quality improvement. To support implementation of this mandate, MINERD has generated school-specific diagnostic reports following each round of the EDN since 2017. These reports benchmark each school's performance against schools serving students with similar socioeconomic backgrounds, as well as against district, regional, and national averages. They also identify subject-specific strengths and weaknesses, with the goal of providing schools with actionable information for developing improvement strategies.

In 2017 and 2018, MINERD began promoting the use of these reports by instructing regional offices and school districts to disseminate the diagnostic results within their jurisdictions through workshops and multiple communication channels. These efforts were strengthened in the 2019–2020 academic year, when MINERD incorporated the use of diagnostic reports into the official school calendar[2] and required schools to complete a sequence of activities based on information from the EDN.

In October 2019, MINERD sent principals an official email assigning them four tasks. First, principals were instructed to retrieve their school-specific reports for the 2017 and 2018 diagnostic assessments through links to the Ministry's online platform. These reports provided the primary input for the subsequent activities. Second, during the week of November 4, schools were ex-

[2]The school calendar is the official annual planning document issued by MINERD and approved by the National Council of Education to regulate teaching and administrative activities in public and private schools in the Dominican Republic. It defines curricular activities, effective instructional weeks, holidays, and evaluation periods across the school year. Distributed before the start of the academic year, it is intended to guide school-level planning and must be used by principals and pedagogical teams in preparing each school's Annual Operating Plan. Thus, activities included in the calendar constitute official mandates that school principals are expected to implement.

pected to organize a workshop with the school leadership team to review the results and identify key strengths and weaknesses in student learning. Third, during the weeks of November 11 and November 18, schools were expected to hold a workshop with teachers to discuss the results and develop improvement plans for the academic year. These plans were intended to guide schools' preparations for the next third- and sixth-grade diagnostic assessments, scheduled for May 2020 and 2021, respectively. Fourth, during the week of November 25, schools were expected to organize a workshop with families to share the results and discuss ways in which parents could support learning at home. Schools were asked to document each workshop with photographs sent through WhatsApp and to email their improvement plans by November 22.

### 2.3 Experimental Design: Sample Selection and Randomized Variation in Reminder Intensity and Content

We designed and implemented the experiment together with MINERD and IDEICE. It targeted primary-school principals and tested whether low-cost reminder calls could reduce the gap between central government instructions and local implementation.

The initial universe comprised 5,386 primary schools that participated in both baseline diagnostic assessments, the third-grade assessment in 2017 and the sixth-grade assessment in 2018. MINERD prepared school-specific diagnostic reports for all schools in this universe and attempted to send the official email communication to their principals.

The experimental sample was restricted to schools satisfying two additional eligibility criteria. First, schools were required to have at least 10 students tested in the sixth-grade diagnostic assessment in 2018, ensuring that baseline achievement measures were not driven by idiosyncratic variation from very small cohorts. A total of 3,811 schools met this requirement. Second, because the intervention relied on direct communication with principals, schools had to have recently validated contact information in IDEICE's records, including both an email address and a phone number. After applying this criterion, the final experimental sample comprised 2,647 schools.[3]

The experimental design was constructed to answer two questions about the implementation of the policy requiring schools to use diagnostic-assessment results. First, does communication intensity matter? Additional reminders may generate larger implementation responses, but their marginal effects may decline once principals have already been contacted. Second, does message content matter? Generic verification calls may be sufficient to attract principals' attention, whereas task-specific reminders may be needed to induce the more substantive actions required by the policy.

To answer these questions, schools in the final sample were randomly assigned within district strata to 12 experimental arms that cross-randomized two dimensions of the intervention: call

[3] Table A1 in the Appendix compares schools in the final experimental sample with schools in the broader diagnostic-assessment universe that were not included in the experiment. The two groups differ on most baseline characteristics, and a joint test rejects equality across groups. Relative to excluded schools, schools in the experimental sample are larger, more likely to be urban and public, and have lower average baseline performance.

intensity and message content. Call intensity ranged from zero to four calls. Message content varied according to the stage of the implementation process targeted by the call. The verification call confirmed whether the principal had received MINERD's email and whether the instructions were clear. The remaining calls reminded principals to organize and report on one of the three workshops required under the policy: a workshop with the school leadership team, a workshop with teachers, and a workshop with families. Four arms were assigned a single call, either the verification call or one of the three workshop reminders, and the other seven combined the verification call with one, two, or three workshop reminders. Schools were assigned to the control group with probability 5/16 and to each of the 11 treatment arms with probability 1/16. The assignment was drawn on October 22, three days before the email was sent. Figure 1 summarizes the design, and Table A2 in the Appendix reports the number of schools assigned to each arm, the number of calls, and their content.

The calls took place in four weekly rounds. Verification calls ran from October 28 to November 1, the week after MINERD sent the email on October 25, and the leadership-team, teacher, and family reminders followed in the weeks of November 4, 11, and 18, as Figure A1 in the Appendix shows. IDEICE's call center placed the calls in the first three rounds, and an external contractor placed those in the fourth.

Table 1 shows that differences in pre-treatment observable characteristics between treatment and control schools are small and, with one exception among the 32 covariates at the 5 percent level, not statistically significant. The joint test fails to reject equality across groups (F-statistic of 0.94, p-value of 0.56). Table A3 in the Appendix extends the comparison to the 12 experimental arms, where a test of equal means across arms rejects at the 5 percent level for one of the 32 covariates. These results indicate that the randomization generated comparable experimental groups.

The follow-up assessments scheduled for 2020 and 2021 were not held, because school closures after the onset of the COVID-19 pandemic in March 2020 suspended the national diagnostic assessments until 2023. We therefore use the third-grade EDN of 2023 and the sixth-grade EDN of 2024 to estimate the longer-run effects of the reminder intervention on student achievement.

# 3 Data Sources and Empirical Strategy

## 3.1 Data Sources

The analysis draws on administrative records from MINERD and original data collected for this study. Together, these sources allow us to measure schools' progression through the policy's implementation chain, from initial engagement with the information campaign to report access, completion of the required activities, preparation of improvement plans, and subsequent student achievement.

Our first data source is the digital platform used by MINERD to distribute the official email communication. The email instructed principals on how to access their school-specific diagnostic

reports and implement the activities required by the policy. The platform records whether each message was opened and whether the recipient clicked on any of the embedded links, as well as the total number of clicks on those links. These outcomes capture the first step in the implementation chain, measuring whether schools were minimally engaged with MINERD's effort to promote the use of diagnostic-assessment information.

Our second data source is MINERD's report-download platform. Rather than being attached directly to the emails, diagnostic reports had to be retrieved from a centralized server by principals or designated school staff. Users were required to access the platform, log in, and download the relevant school-specific reports. The resulting records identify the school associated with each download, the grade and year of the report accessed, and the date and time of each download. These data provide a direct measure of whether schools moved beyond initial engagement with the email communication and actively accessed the information that the policy was designed to disseminate. We measure downloads between October and December 2019 on two margins, whether a school downloaded any report and how many times it did so, counting repeated downloads of the same report.

Our third data source was collected during the campaign to measure implementation of the school-level activities required by the policy. While opening emails, clicking links, and downloading reports were necessary steps, MINERD's objective was for schools to use the results to inform pedagogical planning through the workshops and improvement plans described in Section 2. We use the workshop photographs and the improvement plans that schools submitted to construct measures of compliance with the policy's core implementation requirements.

We also evaluate the content of the submitted improvement plans. Each plan was scored by a large language model (GPT-5.4, accessed through the OpenAI API) that applied a rubric fixed before any outcome or treatment data were examined. The model read the full text of the plan together with the school's own EDN results, which it needs in order to judge whether the plan uses them, and it was never given the school's treatment status. The rubric has seven dimensions. The first records how far the plan uses the results in the school's 2017 and 2018 EDN reports. The next four capture actionability, namely whether the plan establishes specific and realistic goals, proposes concrete and feasible actions, assigns responsibility for individual actions, and specifies dates or implementation deadlines. The last two record whether the plan covers the subjects assessed in the grade and whether the school reflected on how its results informed its decisions. The overall index is the average of the seven dimensions, each rescaled to lie between zero and one. Use of EDN data is an indicator that the plan draws substantively on the school's results rather than merely mentioning them, and actionability is an indicator that the average of the four actionability dimensions is at least one half. Plans are scored separately for grade 3 and grade 6, and each measure is averaged over the school's two grades, with a value of zero for any grade without a plan. A human coder who did not see the model's scores independently scored a stratified sample of plans. Appendix A describes the procedure, the corrections that this validation prompted, and the agreement between the two, which is modest for the use of EDN

data, the dimension that is hardest to judge.

Finally, we measure student achievement using census-based, student-level administrative records from MINERD's EDN. These data include mathematics and Spanish-language scores from the third- and sixth-grade assessments, as well as natural science and social science scores from the sixth-grade assessment. Our post-policy achievement outcomes are the 2023 and 2024 scores. For these assessments we standardize student scores, by grade and subject, to have mean zero and standard deviation one among all students assessed nationally, and we then average them within school, so that the unit of analysis is the school. The 2017 and 2018 scores, which we use as the baseline wave and to assess balance, are standardized in the same way. The 2023 and 2024 assessments were accompanied by questionnaires for principals, teachers and families, which include items on coordination within the school.

### 3.2 Empirical Strategy

Because treatment was randomly assigned within district strata, differences in outcomes across experimental groups identify the causal effects of assignment to reminder calls. We first estimate the effect of reminder calls on compliance with MINERD's dissemination strategy with the following specification:

$$Y_{sd} = \alpha + \sum_{k=1}^{K} \beta_k T_{ks} + \delta_d + \varepsilon_{sd} \tag{1}$$

where $Y_{sd}$ denotes a compliance outcome for school $s$ in district $d$, $\delta_d$ denotes district fixed effects, and $T_{ks}$ is an indicator for assignment to treatment condition $k$. The number of treatment indicators, $K$, varies across specifications. In the pooled specification, $K = 1$, $T_{1s}$ is an indicator equal to one for schools assigned to any reminder-call treatment arm and zero for schools assigned to the no-call control group. The coefficient $\beta_1$ therefore captures the average intent-to-treat (ITT) effect of being assigned to receive reminder calls relative to the control schools, which were sent only the email campaign from MINERD. Two further specifications set $K = 4$. The first groups the 11 treatment arms by the number of calls assigned, from one to four. The second includes one indicator for each message (verification, leadership-team workshop, teacher workshop, family workshop), equal to one if the school was assigned that message. The message indicators are not mutually exclusive, because schools assigned more than one call were assigned more than one message, so each $\beta_k$ is the ITT effect of a message conditional on assignment to the others. In both specifications the omitted group is the no-call control group, and we report tests that the $\beta_k$ are jointly zero and that they are equal to one another. All compliance specifications include district fixed effects to account for the stratified randomization design, and standard errors are clustered at the district level, the randomization stratum.[4] Heteroskedasticity-robust

[4] De Chaisemartin and Ramirez-Cuellar (2024) recommend clustering at the stratum in regressions with stratum fixed effects when strata have ten or fewer units, as 51 of the 121 districts represented in the experimental sample do. Their recommendation is for experiments with many observations per randomized unit. With one observation per unit, as here, they show for paired designs that tests based on heteroskedasticity-robust standard errors at the

standard errors are close to the clustered ones, and of the 90 coefficients in the three compliance tables only two change significance level, both toward greater significance under robust standard errors.

Not every school assigned to a call was reached. Table A5 in the Appendix therefore reports instrumental-variables estimates in which an indicator for at least one delivered call is instrumented with $T_{1s}$. No control school was called, so these estimates are effects of treatment on the treated. We also examine treatment-effect heterogeneity by interacting $T_{1s}$ with an indicator for urban schools or for private schools, and we report these estimates in the Appendix.

We organize the compliance outcomes into two groups corresponding to successive stages of implementation. The first captures initial engagement with MINERD's information campaign, namely whether principals opened the email, whether they clicked on an embedded link, the number of clicks, whether the school downloaded a report, and the number of downloads. The second group captures more demanding actions requiring time and effort from principals and teachers to implement all of the required activities. It includes an indicator for submitting photographic evidence of at least one of the three workshops, an indicator for submitting at least one improvement plan, and the three measures of plan quality described in Section 3.1.

We next examine whether the intervention affected student learning. Exploiting the availability of pre-intervention achievement data, we estimate difference-in-differences specifications on a school panel with two waves per grade, the baseline and the follow-up assessment. The analysis is conducted separately by grade and subject because the diagnostic assessments differ by grade, cohort, and subject coverage.

Our main learning specification estimates the ITT effect of assignment to any reminder-call treatment arm:

$$Y_{sgt} = \alpha + \beta \left(\text{Treatment}_s \times \text{Post}_t\right) + \lambda_s + \text{Post}_t + \varepsilon_{sgt} \tag{2}$$

where $Y_{sgt}$ denotes the standardized achievement outcome for school $s$, grade $g$, and year $t$. $\text{Treatment}_s$ is an indicator equal to one for schools assigned to any reminder-call treatment arm, the $K = 1$ case of $T_{ks}$ in equation (1), and $\text{Post}_t$ is an indicator for the post-intervention assessment period. School fixed effects, $\lambda_s$, absorb time-invariant differences across schools, including baseline achievement levels and other fixed school characteristics. $\text{Post}_t$ captures common changes in achievement between the baseline and post-intervention periods, while $\varepsilon_{sgt}$ denotes the error term, with standard errors clustered at the school level. The coefficient of interest, $\beta$, measures whether achievement changed differentially in treatment relative to control schools after the intervention.

The school panel is unbalanced because 74 schools in grade 3 and 146 schools in grade 6 (147 in natural and social science) are observed in only one wave. These schools remain in the reported sample but do not contribute to $\beta$. Because equation (2) imposes a coefficient of

unit level have correct size as well. Table A4 in the Appendix reports the district-clustered and the heteroskedasticity-robust standard errors side by side for the effects of any reminder call on compliance and on learning.

one on baseline achievement, we also report ANCOVA estimates, which regress the follow-up score on treatment, the baseline score and district fixed effects, with one observation per school and standard errors clustered at the district level.[5] Alongside these estimates we report three diagnostics, namely differential attrition, the baseline gap among schools observed at follow-up, and the persistence of the baseline score.

We further report instrumental-variables estimates of the effect on learning of two implementation actions, reporting at least one of the required workshops and submitting a school improvement plan. Schools that took these actions may differ from other schools in unobserved motivation, managerial capacity, or prior commitment to using assessment data, so we replace $\text{Treatment}_s \times \text{Post}_t$ in equation (2) with each action interacted with $\text{Post}_t$ and use $\text{Treatment}_s \times \text{Post}_t$ as its instrument. These estimates identify a local average treatment effect for schools whose compliance was induced by assignment to reminder calls, not the effect of workshops or improvement plans for all schools. Because the calls moved several actions at once, the exclusion restriction is unlikely to hold for either action taken alone, so we confine these estimates to Table A6 in the Appendix. We also estimate the pooled specification of equation (1) on the coordination indices from the 2023 and 2024 questionnaires, and report the estimates in Table A7 in the Appendix.

# 4 Results

## 4.1 Effects of Any Reminder Call

Table 2 reports the effects of reminder calls on compliance with MINERD's dissemination strategy, and Figure 2 summarizes them in levels, showing for each step of the implementation chain the control-group level and the level implied by the effect of any reminder call. Implementing the workshops and preparing school improvement plans first required that principals engage with the information delivered through the campaign. Although this may appear straightforward, it is far from automatic in many education systems, where messages from the ministry compete with other administrative demands and compliance is not closely monitored.

Panel A shows that, in the absence of reminder calls, engagement with the campaign was far from universal. Even among schools with verified contact information, only 49 percent of principals in the control group opened the campaign email, and only 31 percent clicked on the embedded links. Reminder calls substantially increased both outcomes. Assignment to any treatment arm raised the probability of opening the email by 19 percentage points and that of clicking on the links by 25 percentage points. Reminder calls also increased the intensity of engagement, generating 5.2 additional clicks and 0.7 additional report downloads per school,

[5] To maintain a parsimonious and transparent analysis, the learning regressions focus on the pooled treatment-control contrast rather than the full set of treatment-intensity or message-content arms. Learning is a distal outcome in the implementation chain, and disaggregating treatment arms would substantially increase the number of estimated coefficients without a clear prior that different types of reminders should have differential effects on achievement.

against control means of 5.7 clicks and 1.0 downloads. In addition, only about a third of control-group schools downloaded their school report from MINERD's servers. Assignment to any reminder call increased this probability by 14 percentage points.

Panel B shows that greater engagement translated into concrete school-level action, but also reveals strikingly low levels of observed compliance. Only 11 percent of control schools submitted photographic evidence that at least one workshop had been conducted, and only 5 percent submitted a school improvement plan. This pattern is consistent with the implementation gaps emphasized in the broader literature on policy delivery. Of course, some schools may have carried out these activities without submitting proof to MINERD. Even under that interpretation, however, compliance remains low, since reporting these activities was itself part of the policy. Reminder calls increased the probability of submitting workshop photographs by 27 percentage points and that of submitting an improvement plan by 8 percentage points.

Reminder calls also raised the three measures of plan quality. These are defined over all schools, with a grade for which no plan was submitted counted as zero, so they reflect the additional plans as well as any change in their content. The overall quality index rose by 0.044 relative to a control-group mean of 0.025. The share of a school's two grade-level plans classified as actionable rose by 6.1 percentage points from a control mean of 3.7 percent, and the share making substantive use of the EDN results rose by 2.4 percentage points from a control mean of 0.6 percent. The latter effect is large in relative terms but starts from a very low base.

Taken together, these results show that reminder calls increased both principals' interaction with the campaign and the actions MINERD expected schools to undertake. The effects in Table 2 are ITT effects, and they are close to the effects of the calls themselves. At least one call was delivered to 96 percent of the schools assigned to a call, so the effect of a delivered call, reported in Table A5 in the Appendix, is about 5 percent larger than the corresponding ITT effect. The results also show how far implementation remained from the mandate. As Figure 2 makes clear, even among schools assigned to reminder calls about two thirds opened the email, fewer than half downloaded their report, 38 percent documented a workshop, and 13 percent submitted an improvement plan, of which fewer than a third made substantive use of the EDN results. This low take-up of the mandate is not a concern for the experiment, whose estimates are precise, but it is a concern for education policy more broadly. A mandate that reaches schools through official channels alone is carried out by a small minority of them, and even with reminders most schools stop short of its more demanding tasks.

### 4.2 Effects of Reminder Intensity

We next examine whether outcomes differ with the number of reminder calls assigned. Table 3 reports the results, and Figure 3 plots them for the binary outcomes. Reminder intensity matters across all outcomes, although the pattern differs between initial engagement and subsequent implementation. For early engagement, the evidence points to diminishing marginal returns. One reminder call increases the probability of opening the email by 12 percentage points relative

to the no-call control group and two reminders raise it by 23 percentage points, but the effect then flattens, at 24 percentage points with three reminders and 26 with four. Link clicking follows the same pattern, rising from 15 percentage points with one reminder to 31 with two, with little additional gain from three or four calls, and the effect on report downloads peaks at 20 percentage points with three reminders. The count outcomes in Table 3 are larger with two or more calls than with one, although they do not rise monotonically. The joint tests reject the null of no effect for all engagement outcomes, and the equality tests reject that the coefficients are the same.

The pattern is different for implementation outcomes, where effects continue to increase with reminder intensity. The probability of submitting workshop evidence rises monotonically from 17 percentage points with one reminder to 43 percentage points with four reminders. Likewise, the probability of submitting any improvement plan increases from 5 percentage points with one reminder to 15 percentage points with four reminders. The three plan-quality measures, which count a grade without a plan as zero and so reflect additional plans as well as their content, follow a broadly similar pattern. The overall quality index rises from 0.025 with one reminder to 0.072 with four reminders, an increase driven primarily by gains in actionability (from 0.033 to 0.107) rather than in the use of EDN diagnostic data (from 0.017 to 0.028). Across all five implementation outcomes, the joint tests strongly reject the null of no effect. The equality tests across intensities also reject equal coefficients for four of the five outcomes. The exception is the use of EDN diagnostic data, where the intensity coefficients cannot be statistically distinguished from one another ($p = 0.305$). These results suggest that two reminders may be sufficient to generate most of the gains in initial engagement, but additional reminders remain highly valuable for translating that engagement into reported school-level action and, to a somewhat more limited extent, into higher-quality plans.

### 4.3 Effects of Reminder Content

To study the role of message content, we replace the treatment-intensity indicators with the four message indicators, which enter jointly as described in Section 3.2. Table 4 reports the results, and Figure A2 in the Appendix plots them for the binary outcomes. Across outcomes capturing initial engagement, the verification message is consistently the most effective. Exposure to a verification call increases the probability of opening the email by 15 percentage points and the probability of clicking on the embedded links by 20 percentage points, with an additional 4.8 clicks on average. These effects are somewhat smaller than the average effects of assignment to any reminder call reported above. By contrast, reminders referring to implementing the leadership-team and teacher workshops generate smaller gains in initial engagement, while the family-workshop reminder has no positive effect on clicking behavior or on downloading a report, and is associated with a small negative effect on email opening, compared with the no-call control group.

Results are mixed for report downloads, an initial engagement action that requires somewhat

more effort than simply opening an email or clicking a link. The probability that a school downloaded its report increases only marginally for verification calls but rises by 10 and 9 percentage points for the leadership-team and teacher reminders, respectively. The number of reports downloaded follows a similar pattern, rising most for the leadership-team reminder, while the family-workshop reminder has no detectable effect on either download measure.

The pattern differs for implementation outcomes. Leadership-team and teacher workshop reminders are more strongly associated with concrete school-level action than with initial engagement, while the effects of verification calls become smaller. Message content also maps closely to the type of action being measured. Exposure to the leadership-team reminder increases the probability of submitting workshop photos by 20 percentage points, roughly twice the effect of the verification message and substantially larger than the effects of the other reminders. This pattern is consistent with the idea that the submission of workshop evidence was probably more directly tied to school leadership and coordination teams. By contrast, exposure to the teacher reminder increases the probability of submitting an improvement plan by 9 percentage points, more than twice the effect of the verification message, while the leadership-team message has no detectable effect on plan submission. This is consistent with the view that preparing the improvement plan required greater teacher involvement. The verification message also has positive effects on implementation, but these are smaller. The family-workshop reminder raises workshop-evidence submission by 5 percentage points but has no detectable effect on plan submission.

The same division between the leadership-team and teacher reminders extends to the plan-quality measures. Exposure to the teacher reminder increases the overall quality index by 0.048, roughly twice the control-group mean, with significant gains in both the use of EDN diagnostic data and actionability. The verification message also raises quality, though by less than half as much. Neither the leadership-team nor the family-workshop reminder has a detectable effect on any of the three quality measures, which mirrors the submission results, since these measures count a missing plan as zero and the teacher reminder was the one that moved submission.

The results suggest that different messages were effective at different stages of the process. Verification messages were most useful for generating initial engagement, whereas leadership-team and teacher reminders were more relevant for translating that engagement into implementation. A natural interpretation is that message content matters, but so does timing. As discussed above, verification calls were delivered at the beginning of the campaign, when principals first needed to access the information, whereas workshop-specific reminders were delivered later, in or just before the week in which each workshop was scheduled.

### 4.4 Effects on Student Learning and Within-School Coordination

The intervention was designed to improve student learning by increasing schools' use of diagnostic assessment results. Table 5 reports the effects of reminder calls on student learning under two estimators. These estimates should be interpreted in light of the unusually demanding context in which the intervention operated. They reflect outcomes measured four and five years after the

intervention, after a major disruption to schooling, and for different student cohorts than those directly exposed to the original campaign. Panel A reports the difference-in-differences estimates from equation (2). For third grade, the interaction term is small and statistically insignificant in both mathematics and Spanish. For sixth grade, the estimate indicates a positive effect of about 0.05 standard deviations in mathematics, statistically significant at the 5 percent level, and no effect in Spanish, natural science or social science.

The difference-in-differences specification in equation (2) treats a school's baseline score as a permanent characteristic that the school fixed effect removes, which amounts to imposing a coefficient of one on baseline achievement. Panel B of Table 5 relaxes this restriction with the ANCOVA estimates, and the bottom rows of the table report the three diagnostics. Differential attrition is at most 1.1 percentage points and not statistically significant in any grade or subject. The estimated persistence is about one half to two thirds in every column, well below the value of one that the difference-in-differences imposes, and sixth-grade mathematics is the one outcome in which treated schools start below control schools in the estimation sample, by 0.04 standard deviations. Under these conditions the difference-in-differences attributes to the treatment part of the reversion of lower-scoring treated schools toward the mean. With a persistence of 0.49 and a baseline gap of 0.04 standard deviations, this reversion accounts for roughly half of the 0.05 standard deviation estimate. The ANCOVA estimate for sixth-grade mathematics is 0.025 standard deviations with a standard error of 0.018, positive but not statistically significant, and the two estimators agree on the absence of effects in every other grade and subject. Only one of the six difference-in-differences estimates is statistically significant, and none of the ANCOVA estimates is. We therefore read the sixth-grade mathematics result as suggestive evidence of a learning gain rather than as a precisely established effect.

For completeness, Table A6 in the Appendix reports instrumental-variables estimates that use assignment to reminder calls as an instrument for reporting at least one workshop or for submitting an improvement plan. Because the model is just identified, these estimates are the difference-in-differences estimates rescaled by the first stage, which gives 0.19 standard deviations for workshops and 0.59 for improvement plans in sixth-grade mathematics. They inherit whatever caution applies to the reduced form and to the exclusion restriction discussed in Section 3.2. We therefore do not emphasize them.

Beyond test scores, the EDN questionnaires described in Section 3.1 ask how often the principal meets with teachers and reviews their plans, and how satisfied each party is with that coordination. Table A7 in the Appendix reports the effect of assignment to reminder calls on summary indices of these items and on the two single items that move. Response rates do not differ by treatment status, and the indices are indistinguishable from zero for principals, teachers and families in both years, with 95 percent confidence intervals whose upper bounds do not exceed 0.07. Principals in treated schools report slightly higher satisfaction with their coordination with teachers in both years, by about 0.05 points on a four-point scale, and in 2023 teachers in treated schools are more likely to agree that improvement plans are prepared with

the participation of all actors, a difference that does not reappear in 2024.

### 4.5 Heterogeneous Impacts

Responsiveness to reminder calls may depend on school characteristics, such as connectivity constraints in rural areas or differences in administrative capacity and oversight between public and private schools. Table A8 in the Appendix augments Equation 1, for five of the compliance outcomes of Table 2, with an interaction between the pooled treatment indicator and an indicator for urban schools (Panel A) or for private schools (Panel B). Effects are broadly similar across groups. Urban schools respond somewhat more strongly at the earliest stages of engagement, by an additional 7 percentage points both for opening the email and for downloading the report, significant at the 10 percent level, and no other interaction with location is distinguishable from zero. By sector, engagement does not differ significantly between public and private schools, but private schools show a smaller treatment effect on improvement-plan submission (5 percentage points smaller, significant at the 5 percent level) and on plan quality (0.021 smaller, significant at the 10 percent level). One plausible explanation is that private schools fall outside MINERD's direct administrative hierarchy and face weaker incentives to comply with a ministry-issued mandate once the task becomes demanding, although we cannot test this mechanism with the data at hand. Table A9, also in the Appendix, reports the corresponding estimates for student learning. The difference-in-differences estimate for sixth-grade mathematics is 0.100 standard deviations among rural schools and 0.066 among public schools. No interaction is statistically significant, and these estimates are subject to the same caveat about baseline achievement discussed above. We treat all of this heterogeneity as suggestive.

### 4.6 Cost of the Intervention

We estimate the direct personnel cost of implementing the reminder campaign using IDEICE's call-center payroll for November 2019, the month in which the calls were made. The campaign used IDEICE's existing call-center infrastructure, so its main resource cost was the time of operators and their coordinator. We charge each operator for every day on which the logs record a call by that operator, at the daily equivalent of the operator's salary, and the coordinator for every day with any call. On this basis the four rounds cost 198,331 Dominican pesos, about 3,760 dollars at the November 2019 exchange rate, or 2.07 dollars per school assigned to a call and 0.99 dollars per assigned call (Table A10 in the Appendix). Telephone charges, the email campaign and staff time outside the call center are not included, and the fourth round, which an external contractor placed, is priced at the in-house cost per operator-day. Dividing the cost per treated school by the effects in Table 2 gives the cost of moving one additional school through each step of the mandate, about 11 dollars per additional school that opened the email, 8 dollars per additional school that documented a workshop, and 26 dollars per additional improvement plan, with a 95 percent range of 20 to 35 dollars for the last. The cost per additional plan stays

between 21 and 31 dollars across schools assigned one to four calls, and the cost per additional workshop between 6 and 9 dollars, so heavier reminder schedules deliver implementation at an average price that rises only modestly with the number of calls.

# 5 Conclusions

This paper studies whether low-cost reminders can reduce implementation gaps in education policy in a developing-country context. Embedding an RCT within MINERD's dissemination of diagnostic assessment results, we show that compliance in the absence of additional support was very low. Only a minority of principals opened the official email, clicked on the relevant links, or submitted evidence that schools had carried out the required follow-up actions. Reminder calls substantially increased compliance on all of these margins. Both reminder intensity and message content mattered. Additional calls increased engagement and implementation, although marginal returns diminished for some early engagement outcomes. Message content was most effective when it was aligned with the school actors most directly responsible for the required actions. The message a school received determined which step of the mandate it was more likely to complete. How support is designed shapes whether, and how far, schools translate centrally mandated policies into concrete local action. Direct and repeated communication with principals is, in this sense, a simple implementation technology that a ministry can deploy at scale with the staff and records it already has.

The evidence on whether greater use of standardized assessment results improves student learning is more limited, but still informative. We find a suggestive positive effect in sixth-grade mathematics and no detectable effect elsewhere. The estimate is smaller and imprecise once baseline achievement is modeled rather than differenced out, so we do not regard it as established. At about 2 dollars per school and 26 dollars per additional improvement plan, reminders are an inexpensive way to raise compliance, and any learning gain they produce, if confirmed, would come at that price.

These results should be read alongside two qualifications. First, compliance with the more demanding tasks remained low, and of low quality, even with reminder calls. Only 13 percent of the schools assigned to calls submitted an improvement plan. Nearly all of the plans that were submitted set out concrete actions, but fewer than a third drew substantively on the diagnostic results that the policy was meant to put to use, and we do not observe whether schools implemented these plans in practice. Second, the suggestive achievement gains arise in an unusually demanding setting, with a multi-year gap before follow-up testing, a pandemic-driven disruption to schooling, and a change in student cohorts, conditions that would attenuate almost any true effect. Reminders narrow the gap between policy adoption and implementation, but they do not close it.

More broadly, these results raise three questions for future work. First, how to raise compliance with the most demanding tasks, and the quality of what schools produce, beyond what

reminders achieve. Second, whether interventions that more directly target plan quality and implementation fidelity generate learning gains that are more firmly established than ours and that extend across subjects and grades. Third, whether the gains we document persist as reminder calls lose their novelty and schools' attention returns to competing demands. As assessment systems continue to expand worldwide, the returns to that investment increasingly depend not on generating more information, but on building the institutional capacity to translate the information already produced into concrete, well-implemented action at the school level.

## References


**Agte, Patrick, Daniel R. Morales, Christopher Neilson, Sebastián Otero, and Gautam Rao**, "Incentives, Evidence, and Reminders for Bureaucrats: Overcoming Barriers to Policy Scale Up," NBER Working Paper 35291, National Bureau of Economic Research 2026. Also issued as Yale Economic Growth Center Discussion Paper 1125.

**Andersen, Simon Calmar, Nanna Vestergaard Ahrensberg, Jesper Asring Jessen Hansen, Morten Hjortskov, and Jakob Majlund Holm**, "Attention as a First-Order Barrier to Adoption of Evidence-based Programs," *Journal of Political Economy Microeconomics*, 2026. Forthcoming.

**Andrabi, Tahir, Jishnu Das, and Asim Ijaz Khwaja**, "Report Cards: The Impact of Providing School and Child Test Scores on Educational Markets," *American Economic Review*, 2017, *107* (6), 1535–1563.

**Angrist, Noam and Stefan Dercon**, "Mind the Gap between Education Policy and Practice," *Nature Human Behaviour*, 2024, *8* (12), 2261–2263.

**_ , Tahir Andrabi, Stefan Dercon, Sharnic Djaker, James Habyarimana, Michelle Kaffenberger, Clare Leaver, Akshay Mangla, Nompumelelo Mohohlwane, and Abhijeet Singh**, "An Agenda for Implementation Science in Education," *Science*, 2026, *393* (6816), 1084–1087.

**Antinyan, Armenak, Zareh Asatryan, Zhixin Dai, and Kezhi Wang**, "Does the frequency of reminders matter for their effectiveness? A randomized controlled trial," *Journal of Economic Behavior & Organization*, 2021, *191*, 752–764.

**Banerjee, Abhijit, Arun G. Chandrasekhar, Suresh Dalpath, Esther Duflo, John Floretta, Matthew O. Jackson, Harini Kannan, Francine Loza, Anirudh Sankar, Anna Schrimpf, and Maheshwor Shrestha**, "Selecting the Most Effective Nudge: Evidence from a Large-Scale Experiment on Immunization," *Econometrica*, 2025, *93* (4), 1183–1223.

**Banerjee, Abhijit V., Rukmini Banerji, Esther Duflo, Rachel Glennerster, and Stuti Khemani**, "Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India," *American Economic Journal: Economic Policy*, 2010, *2* (1), 1–30.

**Bergman, Peter and Eric W. Chan**, "Leveraging Parents through Low-Cost Technology: The Impact of High-Frequency Information on Student Achievement," *Journal of Human Resources*, 2021, *56* (1), 125–158.

**Besley, Timothy, Robin Burgess, Adnan Khan, and Guo Xu**, "Bureaucracy and Development," *Annual Review of Economics*, 2022, *14*, 397–424.

**Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur**, "Experimental Evidence on Scaling Up Education Reforms in Kenya," *Journal of Public Economics*, 2018, *168*, 1–20.

**Cheng, Xuejiao and Charles Gale**, "List of National Learning Assessments," Education Policy and Data Center, FHI 360. Archived at https://web.archive.org/web/20140708063710/http://www.epdc.org/education-data-research/list-national-learning-assessments 2014. Original URL discontinued; accessed via the Internet Archive 2026-08-22.

**Cilliers, Jacobus and James Habyarimana**, "Addressing Weak Links in Government Implementation at Scale: Experimental Evidence from a School Governance Reform in Tanzania," *Journal of Development Economics*, 2026, *182*, 103824.

**de Chaisemartin, Clément and Jaime Ramirez-Cuellar**, "At What Level Should One Cluster Standard Errors in Paired and Small-Strata Experiments?," *American Economic Journal: Applied Economics*, 2024, *16* (1), 193–212.

**de Hoyos, Rafael, Alejandro J. Ganimian, and Peter A. Holland**, "Teaching with the Test: Experimental Evidence on Diagnostic Feedback and Capacity Building for Public Schools in Argentina," *The World Bank Economic Review*, 2021, *35* (2), 499–520.

**_ , Sharnic Djaker, Alejandro J. Ganimian, and Peter A. Holland**, "The impact of combining performance-management tools and training with diagnostic feedback in public schools: Experimental evidence from Argentina," *Economics of Education Review*, 2024, *99*, 102518.

**_ , Vicente A. Garcia-Moreno, and Harry Anthony Patrinos**, "The impact of an accountability intervention with diagnostic feedback: Evidence from Mexico," *Economics of Education Review*, 2017, *58*, 123–140.

**Duflo, Annie, Jessica Kiessel, and Adrienne M. Lucas**, "Experimental Evidence on Four Policies to Increase Learning at Scale," *The Economic Journal*, 2024, *134* (661), 1985–2008.

**Dustan, Andrew, Juan Manuel Hernandez-Agramonte, and Stanislao Maldonado**, "Motivating bureaucrats with behavioral insights when state capacity is weak: Evidence from large-scale field experiments in Peru," *Journal of Development Economics*, 2023, *160*, 102995.

**Finan, Frederico, Benjamin A. Olken, and Rohini Pande**, "The Personnel Economics of the Developing State," in Abhijit Vinayak Banerjee and Esther Duflo, eds., *Handbook of Economic Field Experiments*, Vol. 2, North-Holland, 2017, chapter 6, pp. 467–514.

**Karlan, Dean, Margaret McConnell, Sendhil Mullainathan, and Jonathan Zinman**, "Getting to the Top of Mind: How Reminders Increase Saving," *Management Science*, 2016, *62* (12), 3393–3411.

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

**MINERD**, *Anuario de Indicadores Educativos. Año Lectivo 2019-2020*, Santo Domingo: Ministerio de Educación de la República Dominicana, Oficina Nacional de Planificación y Desarrollo Educativo, 2021.

**Muralidharan, Karthik and Abhijeet Singh**, "Improving Public Sector Management at Scale? Experimental Evidence on School Governance in India," NBER Working Paper 28129, National Bureau of Economic Research 2025.

**OECD**, *PISA 2025 Results (Volume I): Future-Ready Students*, Paris: OECD Publishing, 2026.

**Pepinsky, Thomas B., Jan H. Pierskalla, and Audrey Sacks**, "Bureaucracy and Service Delivery," *Annual Review of Political Science*, 2017, *20*, 249–268.

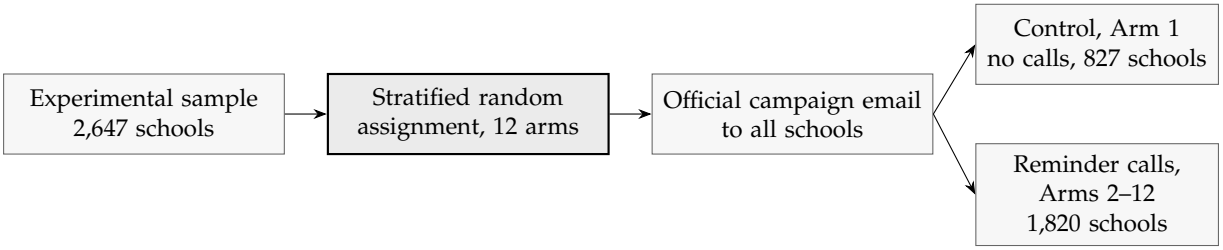
**Rodríguez, Catherine and Juan E. Saavedra**, "The persistent effects of youth savings reminders: Experimental evidence from text-message campaigns in Colombia," *Journal of Development Economics*, 2019, *139*, 135–156.

**UNESCO Institute for Statistics**, "Inventory of Learning Assessments," Montreal: UNESCO Institute for Statistics. https://www.uis.unesco.org/en/data/inventory-learning-assements 2025. Online database covering assessments from 2010 to 2025. The page states no publication date; the year given is the coverage end year. Accessed 2026-08-22. The "assements" spelling is the publisher's own canonical slug.

# Figures

Figure 1: Experimental Design and Treatment Arms

(a) Experimental sequence

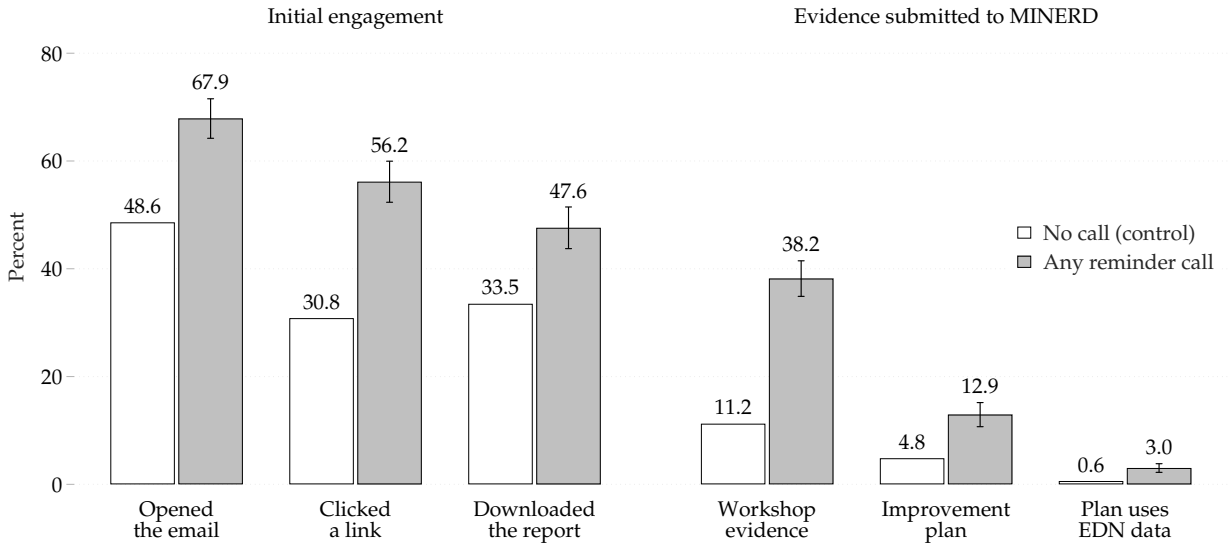



(b) Treatment arms

| | *No call* | *One call* | | | | *Two calls* | | | *Three calls* | | | *Four calls* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Arm** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
| Number of calls | 0 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 3 | 3 | 3 | 4 |
| Verification call | | ● | | | | ● | ● | ● | ● | ● | ● | ● |
| Leadership-team workshop | | | ● | | | ● | | | ● | ● | | ● |
| Teacher workshop | | | | ● | | | ● | | ● | | ● | ● |
| Family workshop | | | | | ● | | | ● | | ● | ● | ● |
| Schools | 827 | 166 | 166 | 165 | 166 | 165 | 166 | 166 | 165 | 165 | 165 | 165 |

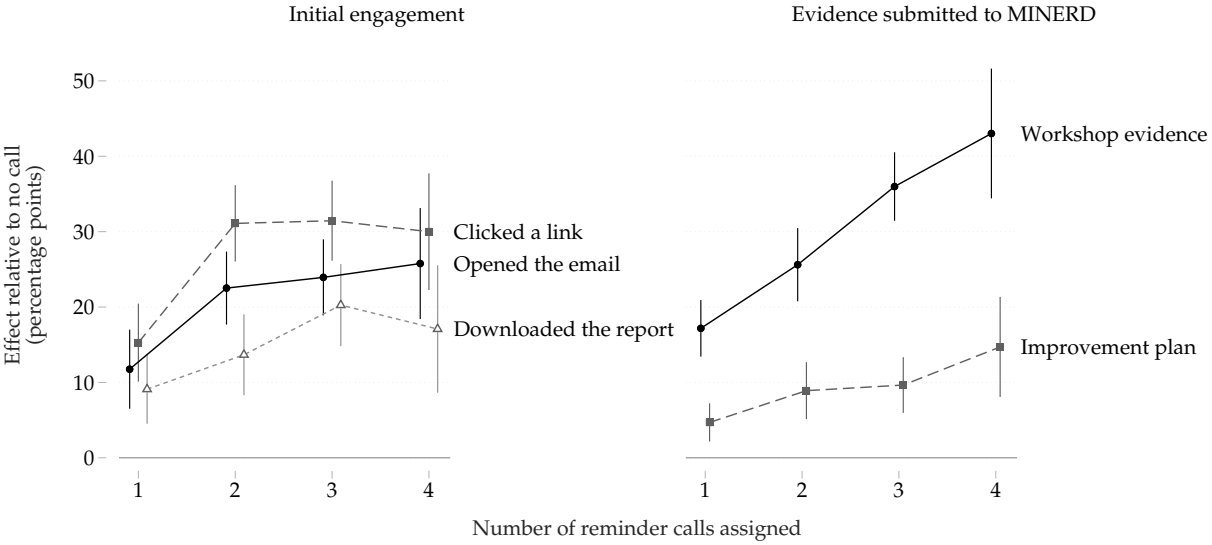
*Notes:* Every school in the experimental sample was sent the official campaign email with the link to its school-specific reports. Arm 1 is the no-call control group. Arms 2 through 12 were additionally assigned the reminder calls marked by a filled circle in panel (b). Schools were assigned to arms at random within district strata. Arm sizes are those of Table A2.

Figure 2: Compliance along the Implementation Chain, with and without Reminder Calls




*Notes:* Each pair of bars refers to one step that MINERD's dissemination strategy asked schools to complete. The open bar is the share of no-call control schools that completed the step. The shaded bar adds the effect of assignment to any reminder call estimated in Table 2, and the vertical line is the 95 percent confidence interval of that effect, based on standard errors clustered at the district level. The numbers above the bars are the bar heights. The steps are not nested, so a school can download its report without having clicked a link in the email. The last step is the share of a school's two grade-level improvement plans that make substantive use of EDN results, with a grade for which no plan was submitted counted as zero. The sample is the 2,647 schools in the experiment.

Figure 3: Effects of Reminder Intensity on Engagement and Implementation




*Notes:* Each marker is the effect of assignment to the indicated number of reminder calls relative to the no-call control group, the zero line, in percentage points, from the regressions reported in Table 3. Vertical lines are 95 percent confidence intervals based on standard errors clustered at the district level. Markers are offset horizontally so that the intervals do not overlap, and the connecting lines are a visual aid. The count outcomes and the measures of plan quality are reported in Table 3.

# Tables

Table 1: Baseline Balance on All Baseline Covariates

| | Control | Treatment | Difference | *p*-value |
|---|---|---|---|---|
| *Panel A. Grade 3 baseline, 2017* | | | | |
| Standardized test score | | | | |
| Mathematics | 0.030 | 0.044 | 0.014 | 0.562 |
| Spanish | 0.025 | 0.043 | 0.018 | 0.459 |
| Share of students at each proficiency level | | | | |
| Spanish, elementary | 0.500 | 0.494 | -0.006 | 0.585 |
| Spanish, acceptable | 0.374 | 0.375 | 0.001 | 0.890 |
| Spanish, satisfactory | 0.126 | 0.131 | 0.005 | 0.440 |
| Mathematics, elementary | 0.440 | 0.433 | -0.007 | 0.513 |
| Mathematics, acceptable | 0.274 | 0.275 | 0.001 | 0.765 |
| Mathematics, satisfactory | 0.286 | 0.291 | 0.005 | 0.582 |
| School characteristics | | | | |
| Number of students tested | 40.213 | 42.102 | 1.889 | 0.175 |
| Public school | 0.728 | 0.712 | -0.016 | 0.402 |
| Full-day school | 0.426 | 0.405 | -0.020 | 0.329 |
| Rural school | 0.391 | 0.374 | -0.017 | 0.405 |
| *Panel B. Grade 6 baseline, 2018* | | | | |
| Standardized test score | | | | |
| Mathematics | 0.081 | 0.044 | -0.037 | 0.155 |
| Spanish | 0.036 | 0.041 | 0.005 | 0.844 |
| Natural science | 0.051 | 0.060 | 0.009 | 0.724 |
| Social science | 0.040 | 0.053 | 0.013 | 0.591 |
| Share of students at each proficiency level | | | | |
| Spanish, elementary | 0.186 | 0.186 | 0.001 | 0.879 |
| Spanish, acceptable | 0.520 | 0.517 | -0.003 | 0.613 |
| Spanish, satisfactory | 0.286 | 0.288 | 0.002 | 0.806 |
| Mathematics, elementary | 0.606 | 0.615 | 0.009 | 0.387 |
| Mathematics, acceptable | 0.331 | 0.332 | 0.001 | 0.931 |
| Mathematics, satisfactory | 0.056 | 0.046 | -0.010 | 0.047 |
| Natural science, elementary | 0.496 | 0.489 | -0.007 | 0.477 |
| Natural science, acceptable | 0.324 | 0.330 | 0.006 | 0.257 |
| Natural science, satisfactory | 0.174 | 0.175 | 0.001 | 0.896 |
| Social science, elementary | 0.390 | 0.385 | -0.004 | 0.612 |
| Social science, acceptable | 0.413 | 0.414 | 0.000 | 0.972 |
| Social science, satisfactory | 0.191 | 0.195 | 0.004 | 0.655 |
| School characteristics | | | | |
| Number of students tested | 41.091 | 42.795 | 1.704 | 0.246 |
| Public school | 0.735 | 0.719 | -0.017 | 0.379 |
| Full-day school | 0.446 | 0.445 | -0.001 | 0.946 |
| Rural school | 0.383 | 0.371 | -0.012 | 0.540 |
| Number of schools | 827 | 1,820 | | |
| Joint *F*-statistic (*p*-value) | | 0.94 (0.560) | | |

*Notes:* This table reports baseline characteristics for schools assigned to any reminder treatment and schools assigned to the no-call control group. The control group is Arm 1, the no-call group, and the treatment group is Arms 2 through 12. Columns 1 and 2 report mean values for control and treated schools, respectively. Column 3 reports the difference in means, treated minus control. Column 4 reports *p*-values from tests of equality of means across the two groups. All variables are measured before random assignment. Treatment was assigned at the school level within district strata. The tests are not clustered by district. A cluster-robust version of the joint test is unreliable with this many restrictions relative to the number of districts. The final row reports the joint test that the baseline covariates are jointly unrelated to treatment assignment. Two of the reported covariates, the satisfactory-level shares for grade-3 Spanish and mathematics, are dropped from that test for collinearity because the three grade-3 proficiency shares sum to one exactly, while the grade-6 shares do not for 19 percent of schools, so it is an F test on 30 of the 32. Every row is on 827 control and 1,820 treated schools except the standardized natural science score, which is on 826 control and 1,820 treated schools because of missing 2018 science scores.

Table 2: Effects of Reminder Calls on Compliance with the Dissemination Strategy

| *Panel A. Initial engagement* | | | | | |
|---|---|---|---|---|---|
| | Email | | | School report | |
| | Opened | Clicked a link | Number of clicks | Downloaded | Number of downloads |
| Treated | 0.193*** | 0.253*** | 5.214*** | 0.141*** | 0.668*** |
| | (0.019) | (0.019) | (0.823) | (0.020) | (0.092) |
| Control mean | 0.486 | 0.308 | 5.677 | 0.335 | 1.034 |
| Observations | 2,647 | 2,647 | 2,647 | 2,647 | 2,647 |
| *Panel B. Evidence submitted to MINERD* | | | | | |
| | Evidence submitted | | Quality of improvement plan | | |
| | Workshop | Improvement plan | Overall index | Use of EDN data | Actionability |
| Treated | 0.269*** | 0.081*** | 0.044*** | 0.024*** | 0.061*** |
| | (0.017) | (0.011) | (0.006) | (0.004) | (0.009) |
| Control mean | 0.112 | 0.048 | 0.025 | 0.006 | 0.037 |
| Observations | 2,647 | 2,647 | 2,647 | 2,647 | 2,647 |

*Notes:* Each coefficient reports the effect of assignment to any reminder-call treatment arm on the indicated outcome relative to the no-call control group. All specifications include district fixed effects. Clustered standard errors at the district level are reported in parentheses. Control means are computed for the no-call control group. Panel A is based on administrative data from the email and report-distribution platforms. Panel B is based on evidence submitted to MINERD and thus captures reported implementation. The three plan-quality measures are school means over the grade 3 and grade 6 plans, with a grade for which no plan was submitted counted as zero, so they reflect additional plans as well as their content. *, **, and *** denote significance at the 10, 5, and 1 percent levels, respectively.

Table 3: Effects of Reminder Intensity on Compliance with the Dissemination Strategy

*Panel A. Initial engagement*

| | Email | | | School report | |
|---|---|---|---|---|---|
| | Opened | Clicked a link | Number of clicks | Downloaded | Number of downloads |
| One call | 0.118*** | 0.153*** | 1.611* | 0.091*** | 0.428*** |
| | (0.027) | (0.026) | (0.858) | (0.023) | (0.120) |
| Two calls | 0.225*** | 0.311*** | 7.309*** | 0.137*** | 0.670*** |
| | (0.024) | (0.026) | (0.954) | (0.027) | (0.126) |
| Three calls | 0.239*** | 0.314*** | 6.675*** | 0.203*** | 0.927*** |
| | (0.025) | (0.027) | (1.011) | (0.027) | (0.143) |
| Four calls | 0.258*** | 0.300*** | 9.049*** | 0.171*** | 0.848*** |
| | (0.037) | (0.039) | (2.185) | (0.043) | (0.262) |
| Control mean | 0.486 | 0.308 | 5.677 | 0.335 | 1.034 |
| Observations | 2,647 | 2,647 | 2,647 | 2,647 | 2,647 |
| Joint $p$-value | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| Equality $p$-value | <0.001 | <0.001 | <0.001 | <0.001 | 0.017 |

*Panel B. Evidence submitted to MINERD*

| | Evidence submitted | | Quality of improvement plan | | |
|---|---|---|---|---|---|
| | Workshop | Improvement plan | Overall index | Use of EDN data | Actionability |
| One call | 0.172*** | 0.047*** | 0.025*** | 0.017*** | 0.033*** |
| | (0.019) | (0.013) | (0.008) | (0.005) | (0.011) |
| Two calls | 0.256*** | 0.089*** | 0.053*** | 0.033*** | 0.074*** |
| | (0.024) | (0.019) | (0.011) | (0.008) | (0.015) |
| Three calls | 0.360*** | 0.097*** | 0.050*** | 0.024*** | 0.071*** |
| | (0.023) | (0.019) | (0.010) | (0.007) | (0.015) |
| Four calls | 0.430*** | 0.147*** | 0.072*** | 0.028** | 0.107*** |
| | (0.043) | (0.034) | (0.017) | (0.012) | (0.026) |
| Control mean | 0.112 | 0.048 | 0.025 | 0.006 | 0.037 |
| Observations | 2,647 | 2,647 | 2,647 | 2,647 | 2,647 |
| Joint $p$-value | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| Equality $p$-value | <0.001 | <0.001 | <0.001 | 0.305 | <0.001 |

*Notes:* Each coefficient reports the effect of assignment to the indicated reminder intensity on the outcome shown relative to the no-call control group. All specifications include district fixed effects. Clustered standard errors at the district level are reported in parentheses. Control means are computed for the no-call control group. Panel A is based on administrative data from the email and report-distribution platforms. Panel B is based on evidence submitted to MINERD and thus captures reported implementation. The three plan-quality measures are school means over the grade 3 and grade 6 plans, with a grade for which no plan was submitted counted as zero, so they reflect additional plans as well as their content. *, **, and *** denote significance at the 10, 5, and 1 percent levels, respectively. The joint $p$-value tests the null that all reminder-intensity coefficients are equal to zero. The $p$-value for equal effects across intensities tests the null that the coefficients for one, two, three, and four calls are equal.

Table 4: Effects of Reminder Content on Compliance with the Dissemination Strategy

*Panel A. Initial engagement*

| | Email | | | School report | |
|---|---|---|---|---|---|
| | Opened | Clicked a link | Number of clicks | Downloaded | Number of downloads |
| Verification call | 0.151*** | 0.202*** | 4.759*** | 0.047* | 0.228* |
| | (0.023) | (0.023) | (0.790) | (0.025) | (0.120) |
| Leadership team workshop call | 0.088*** | 0.080*** | 1.782** | 0.096*** | 0.477*** |
| | (0.021) | (0.019) | (0.821) | (0.020) | (0.121) |
| Teacher workshop call | 0.072*** | 0.059*** | 2.380*** | 0.086*** | 0.300** |
| | (0.022) | (0.021) | (0.846) | (0.027) | (0.121) |
| Family workshop call | -0.044** | -0.004 | -0.107 | 0.000 | 0.078 |
| | (0.021) | (0.020) | (0.733) | (0.022) | (0.108) |
| Control mean | 0.486 | 0.308 | 5.677 | 0.335 | 1.034 |
| Observations | 2,647 | 2,647 | 2,647 | 2,647 | 2,647 |
| Joint *p*-value | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| Equality *p*-value | <0.001 | <0.001 | <0.001 | 0.005 | 0.070 |

*Panel B. Evidence submitted to MINERD*

| | Evidence submitted | | Quality of improvement plan | | |
|---|---|---|---|---|---|
| | Workshop | Improvement plan | Overall index | Use of EDN data | Actionability |
| Verification call | 0.093*** | 0.036** | 0.020** | 0.015** | 0.030** |
| | (0.023) | (0.014) | (0.009) | (0.007) | (0.013) |
| Leadership team workshop call | 0.198*** | 0.019 | 0.007 | -0.006 | 0.011 |
| | (0.024) | (0.014) | (0.009) | (0.007) | (0.012) |
| Teacher workshop call | 0.115*** | 0.088*** | 0.048*** | 0.021*** | 0.067*** |
| | (0.023) | (0.016) | (0.010) | (0.007) | (0.014) |
| Family workshop call | 0.054** | -0.004 | -0.004 | 0.000 | -0.005 |
| | (0.022) | (0.014) | (0.009) | (0.006) | (0.012) |
| Control mean | 0.112 | 0.048 | 0.025 | 0.006 | 0.037 |
| Observations | 2,647 | 2,647 | 2,647 | 2,647 | 2,647 |
| Joint *p*-value | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| Equality *p*-value | <0.001 | <0.001 | <0.001 | 0.041 | <0.001 |

*Notes:* Each coefficient reports the effect of exposure to the indicated reminder call relative to the no-call control group. Since some schools received more than one type of reminder call, the four reminder-content indicators are included jointly in each specification. Coefficients should therefore be interpreted as marginal effects, conditional on exposure to the other call types. All specifications include district fixed effects. Clustered standard errors at the district level are reported in parentheses. Control means are computed for the no-call control group. Panel A is based on administrative data from the email and report-distribution platforms. Panel B is based on evidence submitted to MINERD and thus captures reported implementation. The three plan-quality measures are school means over the grade 3 and grade 6 plans, with a grade for which no plan was submitted counted as zero, so they reflect additional plans as well as their content. The joint *p*-value tests the null that all reminder-content coefficients are equal to zero. The equality *p*-value tests the null that the four reminder-content coefficients are equal. *, **, and *** denote significance at the 10, 5, and 1 percent levels, respectively.

Table 5: Effects of Reminder Calls on Student Learning

| | Grade 3 | | Grade 6 | | | |
|---|---|---|---|---|---|---|
| | Mathematics | Spanish | Mathematics | Spanish | Natural science | Social science |
| *Panel A. Difference-in-differences* | | | | | | |
| Post | 0.052*** | 0.054*** | -0.064*** | -0.017 | -0.036* | -0.024 |
| | (0.018) | (0.018) | (0.020) | (0.015) | (0.019) | (0.016) |
| Treated × Post | -0.005 | -0.016 | 0.051** | -0.001 | 0.005 | 0.008 |
| | (0.022) | (0.022) | (0.024) | (0.019) | (0.023) | (0.020) |
| Schools | 2,647 | 2,647 | 2,647 | 2,647 | 2,647 | 2,647 |
| Observations | 5,220 | 5,220 | 5,148 | 5,148 | 5,147 | 5,147 |
| *Panel B. ANCOVA* | | | | | | |
| Treated | 0.003 | -0.007 | 0.025 | 0.005 | 0.008 | 0.015 |
| | (0.017) | (0.016) | (0.018) | (0.015) | (0.016) | (0.015) |
| Schools | 2,573 | 2,573 | 2,501 | 2,501 | 2,500 | 2,500 |
| *Diagnostics* | | | | | | |
| Differential attrition | -0.011 | -0.011 | -0.006 | -0.006 | -0.006 | -0.005 |
| | (0.006) | (0.006) | (0.008) | (0.008) | (0.008) | (0.008) |
| Control share with follow-up score | 0.979 | 0.979 | 0.949 | 0.949 | 0.949 | 0.948 |
| Baseline gap | 0.012 | 0.018 | -0.042* | 0.010 | 0.006 | 0.014 |
| | (0.026) | (0.026) | (0.023) | (0.022) | (0.024) | (0.024) |
| Persistence of the baseline score | 0.513 | 0.546 | 0.487 | 0.635 | 0.501 | 0.543 |
| Control mean at baseline | 0.030 | 0.025 | 0.081 | 0.036 | 0.051 | 0.040 |
| Control mean at follow-up | 0.083 | 0.080 | 0.020 | 0.013 | 0.014 | 0.014 |

*Notes:* Each column reports the effect of assignment to any reminder call on the standardized test score in the subject shown. For grade 3 the baseline and follow-up assessments are 2017 and 2023, and for grade 6 they are 2018 and 2024. Standard errors are reported in parentheses. Panel A reports difference-in-differences estimates on the school-by-wave panel, with school fixed effects and standard errors clustered at the school level, where Post is an indicator for the follow-up assessment and Treated × Post is the intent-to-treat effect. Panel B reports ANCOVA estimates, in which the follow-up score is regressed on treatment, the baseline score and district fixed effects, with one observation per school observed at follow-up and standard errors clustered at the district level. The diagnostic rows explain why the two panels differ. *Differential attrition* is the treated-minus-control difference in the probability that a school has a follow-up score, on the full experimental sample, with district fixed effects, and the share of control schools with a follow-up score is reported below it. *Baseline gap* is the treated-minus-control difference in the baseline score among the schools observed at follow-up, adjusted for district strata. *Persistence of the baseline score* is the coefficient from regressing the follow-up score on the baseline score. The attrition and baseline-gap regressions cluster standard errors at the district level. The difference-in-differences imposes a persistence of one, so where persistence is below one it converts a share of any baseline gap, one minus persistence times the gap, into an estimated effect of the opposite sign. Control means are for the no-call control group. *, **, and *** denote significance at the 10, 5, and 1 percent levels, respectively.

## Appendix A. Data: Scoring of School Improvement Plans

Schools submitted 439 improvement plans that could be scored, from 279 schools, counting the grade 3 and grade 6 plans of a school separately. The plans arrived in several formats, from MINERD's complete template, with its tables, to free prose. Each plan was scored with a large language model, GPT-5.4 accessed through the OpenAI API, used in the same version throughout. The model received the full text of the plan, including the tables that contain the action plan, together with the school's own results in the 2017 and 2018 EDN. It never received the school's treatment status. The rubric was written and dated before any outcome or treatment data were examined.

Use of the diagnostic results, the first of the seven dimensions, is scored from zero to three and separates plans that merely invoke the school's results from plans that interpret them or derive decisions from them. Goals, actions, responsible parties, deadlines, coverage of the subjects assessed in the grade, and reflection are each scored from zero to two. Reflection refers to the section of the template in which the school explains how its results informed its decisions, and it is left missing when a plan has no such section. Each dimension is rescaled to lie between zero and one. The overall index is their average, the actionability index is the average of goals, actions, responsible parties and deadlines, and a plan is classified as actionable when that index is at least one half. A plan is classified as making substantive use of EDN data when it scores two or more on that first dimension. Among the 439 submitted plans, 96 percent are actionable under this definition and 28 percent make substantive use of the school's results, and the two are only weakly related, with a correlation of about 0.14 between the score for use of the results and the actionability index. Writing a plan with concrete actions was common, and connecting it to the school's results was not.

To validate the scores, an expert human coder scored a stratified sample of plans without seeing the model's scores, and we computed Cohen's kappa between the two. The exercise uncovered two problems, both before any analysis that involved treatment status. First, in nearly all of the plans written on the complete template, the model had not read the tables that contain the action plan and had scored the actionability dimensions and the reflection at zero. These plans were rescored from their full text, after which agreement rose, for example from 0.17 to 0.45 for responsible parties. Second, the model then overstated the use of the diagnostic results, giving the top score to plans that only invoked the results. That dimension was rescored with sharper instructions, and a new validation sample confirmed the improvement. Agreement on the use of the diagnostic results nevertheless remains modest, as the distinction is subtle for human coders as well. Kappa is 0.59 for whether a plan mentions the results at all, 0.37 for the indicator of substantive use that we analyze, and about 0.27 for the full four-point scale, which we therefore do not use. Because the scoring never used treatment status, the error in classifying a given plan is unrelated to assignment. Treated schools submitted more plans, however, so any plan wrongly classified as making substantive use adds to their total, and we read the effect on the use of EDN data with caution.

# Appendix B. Tables

Table A1: Baseline Characteristics of Schools Included and Not Included in the Final RCT Sample

| | Not in final RCT sample | In final RCT sample | Difference | *p*-value |
|---|---|---|---|---|
| *Panel A. Grade 3 baseline, 2017* | | | | |
| Standardized test score | | | | |
| Mathematics | 0.282 | 0.040 | -0.242 | <0.001 |
| Spanish | 0.233 | 0.037 | -0.196 | <0.001 |
| Share of students at each proficiency level | | | | |
| Spanish, elementary | 0.417 | 0.496 | 0.080 | <0.001 |
| Spanish, acceptable | 0.412 | 0.374 | -0.038 | <0.001 |
| Spanish, satisfactory | 0.171 | 0.129 | -0.042 | <0.001 |
| Mathematics, elementary | 0.338 | 0.436 | 0.097 | <0.001 |
| Mathematics, acceptable | 0.280 | 0.275 | -0.005 | 0.220 |
| Mathematics, satisfactory | 0.381 | 0.289 | -0.092 | <0.001 |
| School characteristics | | | | |
| Number of students tested | 17.153 | 41.512 | 24.359 | <0.001 |
| Public school | 0.617 | 0.717 | 0.100 | <0.001 |
| Full-day school | 0.403 | 0.412 | 0.009 | 0.515 |
| Rural school | 0.521 | 0.379 | -0.142 | <0.001 |
| *Panel B. Grade 6 baseline, 2018* | | | | |
| Standardized test score | | | | |
| Mathematics | 0.233 | 0.056 | -0.178 | <0.001 |
| Spanish | 0.086 | 0.039 | -0.046 | 0.006 |
| Natural science | 0.182 | 0.057 | -0.125 | <0.001 |
| Social science | 0.195 | 0.049 | -0.146 | <0.001 |
| Share of students at each proficiency level | | | | |
| Spanish, elementary | 0.175 | 0.186 | 0.011 | 0.018 |
| Spanish, acceptable | 0.511 | 0.518 | 0.007 | 0.242 |
| Spanish, satisfactory | 0.305 | 0.288 | -0.018 | 0.011 |
| Mathematics, elementary | 0.543 | 0.612 | 0.069 | <0.001 |
| Mathematics, acceptable | 0.376 | 0.331 | -0.045 | <0.001 |
| Mathematics, satisfactory | 0.074 | 0.049 | -0.025 | <0.001 |
| Natural science, elementary | 0.443 | 0.491 | 0.048 | <0.001 |
| Natural science, acceptable | 0.338 | 0.328 | -0.010 | 0.049 |
| Natural science, satisfactory | 0.213 | 0.174 | -0.039 | <0.001 |
| Social science, elementary | 0.345 | 0.387 | 0.041 | <0.001 |
| Social science, acceptable | 0.400 | 0.413 | 0.014 | 0.008 |
| Social science, satisfactory | 0.249 | 0.194 | -0.055 | <0.001 |
| School characteristics | | | | |
| Number of students tested | 16.641 | 42.262 | 25.621 | <0.001 |
| Public school | 0.619 | 0.724 | 0.105 | <0.001 |
| Full-day school | 0.429 | 0.445 | 0.016 | 0.241 |
| Rural school | 0.518 | 0.375 | -0.144 | <0.001 |
| Number of schools | 2,739 | 2,647 | | |
| Joint *F*-statistic (*p*-value) | | 38.24 (<0.001) | | |

*Notes:* This table compares baseline characteristics of schools included in the final experimental sample with schools in the broader diagnostic-assessment universe that were not included in the experiment. The universe includes primary schools that administered both the grade 3 diagnostic assessment in 2017 and the grade 6 diagnostic assessment in 2018. The final RCT sample is restricted to schools with at least 10 students tested in grade 6 in 2018 and recently validated principal contact information, including both an email address and a phone number. Columns 1 and 2 report the mean for schools not included in and included in the final RCT sample, respectively. Column 3 is column 2 minus column 1 on unrounded means, so it may differ in the last digit from the difference of the rounded columns. The *p*-value is from a two-sided test of equality of means. The school counts apply to all rows except the standardized grade-6 scores, on 2,737 excluded and 2,647 included schools, and the natural science score, on 2,734 and 2,646. The final row is the joint F test, from a regression of the inclusion indicator on all covariates with robust standard errors, that the covariates are jointly unrelated to inclusion. The two grade-3 satisfactory shares drop for collinearity, so it is a test on 30 of the 32.

Table A2: Experimental Design and Treatment Assignment

| Arm | Reminder calls assigned | Schools | Percent |
|---|---|---|---|
| *No call* | | | |
| 1 | None (control) | 827 | 31.2 |
| *One call* | | | |
| 2 | Verification | 166 | 6.3 |
| 3 | Leadership team workshop | 166 | 6.3 |
| 4 | Teacher workshop | 165 | 6.2 |
| 5 | Family workshop | 166 | 6.3 |
| *Two calls* | | | |
| 6 | Verification, leadership team workshop | 165 | 6.2 |
| 7 | Verification, teacher workshop | 166 | 6.3 |
| 8 | Verification, family workshop | 166 | 6.3 |
| *Three calls* | | | |
| 9 | Verification, leadership team, teacher workshops | 165 | 6.2 |
| 10 | Verification, leadership team, family workshops | 165 | 6.2 |
| 11 | Verification, teacher, family workshops | 165 | 6.2 |
| *Four calls* | | | |
| 12 | Verification, leadership team, teacher, family workshops | 165 | 6.2 |
| | Total | 2,647 | 100.0 |

*Notes:* This table reports the experimental design and treatment assignment. Schools were randomly assigned within district strata to one of 12 experimental arms that varied in both the number of calls assigned and the content of those calls. Arm 1 is the no-call control group and Arms 2 through 12 are the treatment arms. The verification call confirmed whether the principal had received the email sent by MINERD and whether the instructions were clear. Workshop reminder calls encouraged principals to organize and report on the workshops with the leadership team, teachers, or families, following MINERD's implementation schedule. Percentages are shares of the 2,647 schools in the experimental sample and do not sum exactly to 100 because of rounding.

Table A3: Baseline Balance Across Experimental Arms

| | Treatment arm | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | *p*-value |
| *Panel A. Grade 3 baseline, 2017* | | | | | | | | | | | | | |
| Standardized test score | | | | | | | | | | | | | |
| Mathematics | 0.03 | 0.07 | 0.07 | 0.01 | -0.01 | 0.02 | 0.04 | 0.12 | 0.04 | 0.06 | 0.03 | 0.04 | 0.869 |
| Spanish | 0.02 | 0.08 | 0.08 | 0.01 | 0.03 | 0.04 | 0.01 | 0.10 | 0.02 | 0.05 | 0.00 | 0.06 | 0.871 |
| Share of students at each proficiency level | | | | | | | | | | | | | |
| Spanish, elementary | 0.50 | 0.48 | 0.48 | 0.51 | 0.51 | 0.50 | 0.52 | 0.46 | 0.50 | 0.48 | 0.51 | 0.48 | 0.698 |
| Spanish, acceptable | 0.37 | 0.37 | 0.38 | 0.35 | 0.37 | 0.37 | 0.36 | 0.39 | 0.38 | 0.38 | 0.37 | 0.40 | 0.212 |
| Spanish, satisfactory | 0.13 | 0.15 | 0.14 | 0.14 | 0.12 | 0.13 | 0.13 | 0.14 | 0.12 | 0.13 | 0.12 | 0.12 | 0.536 |
| Mathematics, elementary | 0.44 | 0.43 | 0.42 | 0.45 | 0.46 | 0.44 | 0.44 | 0.41 | 0.44 | 0.43 | 0.44 | 0.43 | 0.939 |
| Mathematics, acceptable | 0.27 | 0.27 | 0.27 | 0.27 | 0.28 | 0.29 | 0.26 | 0.26 | 0.27 | 0.27 | 0.28 | 0.30 | 0.119 |
| Mathematics, satisfactory | 0.29 | 0.31 | 0.30 | 0.28 | 0.26 | 0.28 | 0.30 | 0.33 | 0.29 | 0.30 | 0.28 | 0.27 | 0.396 |
| School characteristics | | | | | | | | | | | | | |
| Number of students tested | 40.21 | 41.99 | 39.68 | 42.17 | 44.57 | 42.82 | 42.52 | 39.35 | 40.95 | 41.79 | 46.34 | 40.96 | 0.815 |
| Public school | 0.73 | 0.67 | 0.73 | 0.68 | 0.77 | 0.68 | 0.73 | 0.68 | 0.71 | 0.71 | 0.72 | 0.75 | 0.618 |
| Full-day school | 0.43 | 0.42 | 0.40 | 0.36 | 0.43 | 0.40 | 0.43 | 0.41 | 0.41 | 0.38 | 0.42 | 0.39 | 0.962 |
| Rural school | 0.39 | 0.33 | 0.42 | 0.43 | 0.34 | 0.37 | 0.34 | 0.36 | 0.36 | 0.39 | 0.38 | 0.38 | 0.704 |
| *Panel B. Grade 6 baseline, 2018* | | | | | | | | | | | | | |
| Standardized test score | | | | | | | | | | | | | |
| Mathematics | 0.08 | 0.08 | 0.03 | 0.06 | -0.02 | 0.02 | 0.04 | 0.09 | 0.06 | 0.03 | 0.06 | 0.04 | 0.849 |
| Spanish | 0.04 | 0.12 | 0.07 | 0.03 | 0.02 | 0.02 | 0.03 | 0.04 | 0.06 | 0.01 | 0.02 | 0.02 | 0.925 |
| Natural science | 0.05 | 0.13 | 0.02 | 0.00 | 0.02 | 0.07 | 0.06 | 0.12 | 0.07 | 0.06 | 0.05 | 0.05 | 0.745 |
| Social science | 0.04 | 0.14 | 0.03 | 0.08 | 0.00 | 0.03 | 0.05 | 0.08 | 0.08 | 0.02 | 0.03 | 0.03 | 0.706 |
| Share of students at each proficiency level | | | | | | | | | | | | | |
| Spanish, elementary | 0.19 | 0.17 | 0.19 | 0.20 | 0.18 | 0.19 | 0.19 | 0.19 | 0.18 | 0.20 | 0.19 | 0.19 | 0.922 |
| Spanish, acceptable | 0.52 | 0.51 | 0.49 | 0.50 | 0.54 | 0.52 | 0.53 | 0.51 | 0.52 | 0.51 | 0.53 | 0.52 | 0.409 |
| Spanish, satisfactory | 0.29 | 0.31 | 0.31 | 0.29 | 0.27 | 0.28 | 0.28 | 0.30 | 0.30 | 0.28 | 0.28 | 0.28 | 0.815 |
| Mathematics, elementary | 0.61 | 0.59 | 0.63 | 0.61 | 0.63 | 0.62 | 0.62 | 0.60 | 0.61 | 0.62 | 0.61 | 0.62 | 0.947 |
| Mathematics, acceptable | 0.33 | 0.36 | 0.33 | 0.33 | 0.32 | 0.32 | 0.33 | 0.33 | 0.34 | 0.33 | 0.34 | 0.33 | 0.910 |
| Mathematics, satisfactory | 0.06 | 0.04 | 0.04 | 0.05 | 0.04 | 0.05 | 0.05 | 0.06 | 0.04 | 0.04 | 0.04 | 0.05 | 0.765 |
| Natural science, elementary | 0.50 | 0.46 | 0.49 | 0.52 | 0.50 | 0.48 | 0.50 | 0.48 | 0.48 | 0.48 | 0.49 | 0.49 | 0.700 |
| Natural science, acceptable | 0.32 | 0.34 | 0.34 | 0.31 | 0.33 | 0.34 | 0.31 | 0.32 | 0.34 | 0.34 | 0.34 | 0.34 | 0.035 |
| Natural science, satisfactory | 0.17 | 0.20 | 0.16 | 0.17 | 0.16 | 0.18 | 0.18 | 0.20 | 0.17 | 0.17 | 0.17 | 0.17 | 0.737 |
| Social science, elementary | 0.39 | 0.36 | 0.40 | 0.39 | 0.40 | 0.39 | 0.39 | 0.37 | 0.36 | 0.40 | 0.40 | 0.39 | 0.463 |
| Social science, acceptable | 0.41 | 0.42 | 0.40 | 0.39 | 0.42 | 0.43 | 0.41 | 0.43 | 0.43 | 0.40 | 0.41 | 0.42 | 0.193 |
| Social science, satisfactory | 0.19 | 0.22 | 0.19 | 0.21 | 0.18 | 0.18 | 0.19 | 0.20 | 0.20 | 0.19 | 0.19 | 0.19 | 0.693 |
| School characteristics | | | | | | | | | | | | | |
| Number of students tested | 41.09 | 40.57 | 39.65 | 42.62 | 45.69 | 42.85 | 43.20 | 40.04 | 43.39 | 43.19 | 47.30 | 42.26 | 0.777 |
| Public school | 0.74 | 0.68 | 0.73 | 0.68 | 0.78 | 0.70 | 0.74 | 0.69 | 0.73 | 0.72 | 0.72 | 0.75 | 0.657 |
| Full-day school | 0.45 | 0.46 | 0.45 | 0.41 | 0.46 | 0.44 | 0.50 | 0.43 | 0.44 | 0.42 | 0.45 | 0.43 | 0.973 |
| Rural school | 0.38 | 0.33 | 0.42 | 0.43 | 0.34 | 0.37 | 0.34 | 0.36 | 0.36 | 0.39 | 0.37 | 0.37 | 0.748 |
| Number of schools | 827 | 166 | 166 | 165 | 166 | 165 | 166 | 166 | 165 | 165 | 165 | 165 | |

*Notes:* This table reports baseline characteristics by randomized cell. Columns 1 to 12 report mean values for each experimental arm, numbered as in Table A2, with Arm 1 the no-call control group. The final column reports the *p*-value from a test of equality of means across the twelve randomized cells for each baseline variable, obtained by regressing the covariate on the full set of arm indicators and testing them jointly. All variables are measured before random assignment. Treatment was assigned at the school level within district strata.

Table A4: Clustered and Robust Standard Errors for the Main Estimates

| | Coefficient | Clustered standard error | Robust standard error | Ratio | Schools |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| *Panel A. Initial engagement* | | | | | |
| Opened the email | 0.193 | (0.019)*** | (0.021)*** | 0.90 | 2,647 |
| Clicked a link | 0.253 | (0.019)*** | (0.020)*** | 0.98 | 2,647 |
| Number of clicks | 5.214 | (0.823)*** | (0.819)*** | 1.00 | 2,647 |
| Downloaded the report | 0.141 | (0.020)*** | (0.019)*** | 1.04 | 2,647 |
| Number of downloads | 0.668 | (0.092)*** | (0.098)*** | 0.94 | 2,647 |
| *Panel B. Evidence submitted to MINERD* | | | | | |
| Workshop | 0.269 | (0.017)*** | (0.016)*** | 1.06 | 2,647 |
| Improvement plan | 0.081 | (0.011)*** | (0.011)*** | 1.03 | 2,647 |
| Overall index of the plan | 0.044 | (0.006)*** | (0.006)*** | 1.01 | 2,647 |
| Plan uses EDN data | 0.024 | (0.004)*** | (0.004)*** | 1.01 | 2,647 |
| Plan is actionable | 0.061 | (0.009)*** | (0.009)*** | 1.03 | 2,647 |
| *Panel C. Student learning, ANCOVA* | | | | | |
| Grade 3 | | | | | |
| Mathematics | 0.003 | (0.017) | (0.018) | 0.96 | 2,573 |
| Spanish | -0.007 | (0.016) | (0.017) | 0.98 | 2,573 |
| Grade 6 | | | | | |
| Mathematics | 0.025 | (0.018) | (0.018) | 0.99 | 2,501 |
| Spanish | 0.005 | (0.015) | (0.016) | 0.94 | 2,501 |
| Natural science | 0.008 | (0.016) | (0.017) | 0.95 | 2,500 |
| Social science | 0.015 | (0.015) | (0.016) | 0.92 | 2,500 |

*Notes:* Each row reports one coefficient of assignment to any reminder call, first with the standard error clustered at the district level, which is what Table 2 and Panel B of Table 5 print, and then with a heteroskedasticity-robust standard error that treats each school as independent. Stars are placed on each standard error and come from the test that uses it, with the t distribution and the degrees of freedom of that variance estimator, so a difference in stars between the two columns marks a coefficient whose significance level depends on the choice. *Ratio* is the clustered standard error divided by the robust one. Panels A and B are the ten outcomes of Table 2, estimated on the full experimental sample with district fixed effects. Panel C is the ANCOVA of Table 5, in which the follow-up score is regressed on treatment, the baseline score and district fixed effects on the schools observed at follow-up. Treatment was assigned within district strata, and 51 of the 121 districts have ten or fewer schools. Over the 90 coefficients of Tables 2, 3 and 4 the ratio ranges from 0.85 to 1.25, and two coefficients change significance level, both toward greater significance under robust standard errors, none of them among the coefficients printed here. In Panel C the ratio ranges from 0.92 to 0.99 and no coefficient changes significance level. *, **, and *** denote significance at the 10, 5, and 1 percent levels, respectively.

Table A5: Effects of a Delivered Reminder Call on Compliance with the Dissemination Strategy

| *Panel A. Initial engagement* | | | | | |
|---|---|---|---|---|---|
| | Email | | | School report | |
| | Opened | Clicked a link | Number of clicks | Downloaded | Number of downloads |
| Call delivered | 0.201*** | 0.265*** | 5.448*** | 0.147*** | 0.698*** |
| | (0.019) | (0.020) | (0.854) | (0.020) | (0.096) |
| Control mean | 0.486 | 0.308 | 5.677 | 0.335 | 1.034 |
| Observations | 2,647 | 2,647 | 2,647 | 2,647 | 2,647 |
| *Panel B. Evidence submitted to MINERD* | | | | | |
| | Evidence submitted | | Quality of improvement plan | | |
| | Workshop | Improvement plan | Overall index | Use of EDN data | Actionability |
| Call delivered | 0.281*** | 0.085*** | 0.046*** | 0.025*** | 0.064*** |
| | (0.017) | (0.012) | (0.006) | (0.004) | (0.010) |
| Control mean | 0.112 | 0.048 | 0.025 | 0.006 | 0.037 |
| Observations | 2,647 | 2,647 | 2,647 | 2,647 | 2,647 |

*Notes:* Each coefficient is an instrumental-variables estimate of the effect of having at least one reminder call delivered, meaning that the caller transmitted the full message to the school, instrumented with assignment to any reminder-call treatment arm. Of the 1,820 schools assigned to a call, 1,742 had at least one call delivered, and no control school was called, so the estimates are effects of treatment on the treated. School-rounds with no recorded delivery outcome (162 of 3,802) are counted as not delivered. If every school with an unrecorded round had in fact been reached, the first stage would be 0.985 and each estimate about 2.8 percent smaller. The first-stage coefficient on assignment is 0.957 (standard error 0.005), with a Kleibergen–Paap excluded-instrument F of 39,998, and it is common to all columns because they share one sample. Each estimate therefore equals the corresponding intent-to-treat effect in Table 2 divided by the first stage. All specifications include district fixed effects. Clustered standard errors at the district level are reported in parentheses. Control means are computed for the no-call control group. The three plan-quality measures are school means over the grade 3 and grade 6 plans, with a grade for which no plan was submitted counted as zero, so they reflect additional plans as well as their content. *, **, and *** denote significance at the 10, 5, and 1 percent levels, respectively.

Table A6: Instrumental-Variables Estimates of the Effects of Compliance on Student Learning

| | Grade 3 | | Grade 6 | | | |
|---|---|---|---|---|---|---|
| | Mathematics | Spanish | Mathematics | Spanish | Natural science | Social science |
| *Panel A. Reduced form, effect of any reminder call* | | | | | | |
| Post | 0.052*** | 0.054*** | -0.064*** | -0.017 | -0.036* | -0.024 |
| | (0.018) | (0.018) | (0.020) | (0.015) | (0.019) | (0.016) |
| Treated × Post | -0.005 | -0.016 | 0.051** | -0.001 | 0.005 | 0.008 |
| | (0.022) | (0.022) | (0.024) | (0.019) | (0.023) | (0.020) |
| *Panel B. Effect of reporting a workshop* | | | | | | |
| Post | 0.054** | 0.061** | -0.086*** | -0.016 | -0.038 | -0.027 |
| | (0.027) | (0.026) | (0.029) | (0.022) | (0.027) | (0.023) |
| Workshop × Post | -0.020 | -0.060 | 0.186** | -0.004 | 0.019 | 0.029 |
| | (0.081) | (0.079) | (0.089) | (0.068) | (0.082) | (0.074) |
| *First stage* | | | | | | |
| Treated × Post | 0.273*** | 0.273*** | 0.276*** | 0.276*** | 0.276*** | 0.276*** |
| | (0.016) | (0.016) | (0.016) | (0.016) | (0.016) | (0.016) |
| Kleibergen–Paap F | 288.8 | 288.8 | 285.3 | 285.3 | 284.7 | 284.7 |
| *Panel C. Effect of submitting an improvement plan* | | | | | | |
| Post | 0.055* | 0.064** | -0.093*** | -0.016 | -0.039 | -0.028 |
| | (0.030) | (0.029) | (0.033) | (0.025) | (0.030) | (0.026) |
| Improvement plan × Post | -0.065 | -0.196 | 0.592** | -0.014 | 0.062 | 0.093 |
| | (0.265) | (0.258) | (0.290) | (0.217) | (0.261) | (0.234) |
| *First stage* | | | | | | |
| Treated × Post | 0.084*** | 0.084*** | 0.087*** | 0.087*** | 0.087*** | 0.087*** |
| | (0.011) | (0.011) | (0.011) | (0.011) | (0.011) | (0.011) |
| Kleibergen–Paap F | 57.0 | 57.0 | 59.4 | 59.4 | 59.2 | 59.2 |
| Schools | 2,647 | 2,647 | 2,647 | 2,647 | 2,647 | 2,647 |
| Observations | 5,220 | 5,220 | 5,148 | 5,148 | 5,147 | 5,147 |

*Notes:* Each column reports estimates from the school-by-wave panel used in Panel A of Table 5, with school fixed effects. Standard errors clustered at the school level are reported in parentheses. Post is an indicator for the follow-up assessment. The dependent variable is the standardized test score in the subject shown. Panel A reproduces the reduced form, the difference-in-differences effect of assignment to any reminder call. Panels B and C report just-identified instrumental-variables estimates, where Workshop × Post and Improvement plan × Post are each instrumented with Treated × Post. Because the model is just identified, each instrumental-variables estimate is the Panel A reduced form divided by that column's first stage, so Panels B and C carry no information beyond Panel A and inherit the caveats discussed in the text. They also require that assignment affects learning only through the instrumented action, which is unlikely to hold for either action taken alone since the calls moved both. The first stage is reported for every column, and the F-statistic is the Kleibergen–Paap excluded-instrument F. *, **, and *** denote significance at the 10, 5, and 1 percent levels, respectively.

Table A7: Effects of Reminder Calls on Within-School Coordination Reported in the Assessment Questionnaires

| | Director questionnaire | | Teacher questionnaire | | Family questionnaire |
|---|---|---|---|---|---|
| | Index | Satisfaction with coordination with teachers | Index | Improvement plans involve all actors | Index |
| | (1) | (2) | (3) | (4) | (5) |
| *Panel A. 2023 questionnaires, grade 3* | | | | | |
| Treated | 0.020 | 0.047** | 0.005 | 0.055** | 0.002 |
| | (0.025) | (0.024) | (0.024) | (0.023) | (0.010) |
| Control mean | 0.001 | 3.429 | 0.002 | 3.291 | 0.013 |
| Schools | 2,178 | 2,178 | 2,118 | 2,118 | 2,504 |
| *Panel B. 2024 questionnaires, grade 6* | | | | | |
| Treated | 0.011 | 0.053* | -0.013 | -0.011 | -0.011 |
| | (0.022) | (0.031) | (0.027) | (0.026) | (0.010) |
| Control mean | -0.002 | 3.365 | 0.017 | 3.359 | 0.000 |
| Schools | 1,943 | 1,943 | 1,722 | 1,722 | 2,448 |

*Notes:* Each column reports the effect of the reminder calls on a measure of within-school coordination taken from the questionnaires that directors, teachers and families answered alongside the 2023 (grade 3, Panel A) and 2024 (grade 6, Panel B) national diagnostic assessments, four and five years after the calls. There is one observation per school, and the teacher and family measures are school means of the respondents. *Treated* is the intent-to-treat estimate of assignment to any reminder call, with district fixed effects and standard errors clustered at the district level. *Control mean* is the mean of the outcome in the no-call control schools in the estimation sample, and *Schools* is the number of schools with a response. Indices are control-group standardized means, with higher values indicating more coordination, of the following items. The director index covers the frequency of meetings with teachers, of meetings of the pedagogical group, of reviewing teachers' lesson plans, of analyzing student assessment reports, and of following up on agreements with teachers or district staff, satisfaction with the relationship and coordination with teaching staff, poor coordination between teachers and management not being a problem, and, after the assessment results, having met with the grade's teachers, held a workshop with the cycle's teachers, and drawn up an action plan involving everyone. The teacher index covers satisfaction with the support received for lesson planning, agreement that management stays informed of student performance, makes improvement plans involving all actors, keeps track of teachers' work, and gives effective suggestions on teaching practice, and the frequency with which the coordinator or director reviewed the teacher's lesson plans, reviewed the teacher's student assessments, observed and gave feedback on classes, and gave materials and guidance after a diagnosis, and with which the teacher attended meetings of teachers with management and meetings of teachers by cycle or grade. The family index covers attendance at meetings organized by the teacher and at meetings to receive the student's evaluation report, and satisfaction with the school's support to parents and with the information it gives on the student's progress. The director satisfaction column and the teacher improvement-plans column are single items on a four-point scale. Response rates do not differ between treated and control schools in any questionnaire or year. *, **, and *** denote significance at the 10, 5, and 1 percent levels, respectively.

Table A8: Heterogeneous Effects of Reminder Calls by School Location and Sector

| | Opened email | Downloaded report | Workshop evidence | Improvement plan | Overall index |
|---|---|---|---|---|---|
| *Panel A. By school location* | | | | | |
| Treated | 0.148*** | 0.099*** | 0.281*** | 0.071*** | 0.044*** |
| | (0.032) | (0.030) | (0.025) | (0.018) | (0.010) |
| Treated × Urban | 0.071* | 0.067* | -0.018 | 0.016 | -0.001 |
| | (0.039) | (0.036) | (0.030) | (0.023) | (0.012) |
| *Panel B. By school sector* | | | | | |
| Treated | 0.186*** | 0.127*** | 0.282*** | 0.095*** | 0.050*** |
| | (0.021) | (0.021) | (0.020) | (0.013) | (0.007) |
| Treated × Private | 0.023 | 0.057 | -0.043 | -0.047** | -0.021* |
| | (0.043) | (0.047) | (0.029) | (0.020) | (0.011) |
| Control mean | 0.486 | 0.335 | 0.112 | 0.048 | 0.025 |
| Observations | 2,647 | 2,647 | 2,647 | 2,647 | 2,647 |

*Notes:* Each panel presents the heterogeneous impacts of the pooled reminder-call treatment indicator and the subgroup indicator shown. Panel A interacts treatment with an indicator for urban schools, so the omitted category is rural schools. Panel B interacts treatment with an indicator for private schools, so the omitted category is public schools. Outcomes are the compliance and plan-quality measures of Table 2. All specifications include district fixed effects, and standard errors are clustered at the district level. The rows report the treatment indicator and its interaction with the subgroup. Heterogeneous effects on student learning are reported in Table A9. Control means are computed for the no-call control group. *, **, and *** denote significance at the 10, 5, and 1 percent levels, respectively.

Table A9: Heterogeneous Effects of Reminder Calls on Student Learning by School Location and Sector

| | Grade 3 | | Grade 6 | | | |
|---|---|---|---|---|---|---|
| | Mathematics | Spanish | Mathematics | Spanish | Natural science | Social science |
| *Panel A. By school location* | | | | | | |
| Treated × Post | -0.031 | -0.017 | 0.100** | 0.003 | 0.028 | 0.015 |
| | (0.042) | (0.040) | (0.047) | (0.035) | (0.045) | (0.039) |
| Treated × Post × Urban | 0.042 | -0.001 | -0.081 | -0.004 | -0.038 | -0.011 |
| | (0.049) | (0.047) | (0.054) | (0.041) | (0.051) | (0.045) |
| *Panel B. By school sector* | | | | | | |
| Treated × Post | -0.014 | -0.026 | 0.066** | 0.005 | 0.011 | 0.008 |
| | (0.027) | (0.026) | (0.029) | (0.022) | (0.027) | (0.024) |
| Treated × Post × Private | 0.034 | 0.034 | -0.056 | -0.015 | -0.020 | 0.005 |
| | (0.047) | (0.044) | (0.054) | (0.040) | (0.048) | (0.044) |
| Control mean at baseline | 0.030 | 0.025 | 0.081 | 0.036 | 0.051 | 0.040 |
| Observations | 5,220 | 5,220 | 5,148 | 5,148 | 5,147 | 5,147 |

*Notes:* Each column reports estimates from the school-level difference-in-differences specification of Panel A of Table 5, augmented with an interaction between the subgroup indicator and every term. The dependent variable is the standardized test score in the subject indicated. For grade 3 the baseline and follow-up assessments are 2017 and 2023, and for grade 6 they are 2018 and 2024. Panel A interacts with an indicator for urban schools, so the omitted category is rural schools. Panel B interacts with an indicator for private schools, so the omitted category is public schools. Treated × Post is the effect of assignment to reminder calls in the omitted category, and the interaction row is the differential effect in the subgroup shown. All specifications include school fixed effects, and standard errors are clustered at the school level. Control means are the baseline mean of the no-call control group. *, **, and *** denote significance at the 10, 5, and 1 percent levels, respectively.

Table A10: Cost and Cost-Effectiveness of the Reminder Campaign

*Panel A. Cost of the campaign*

| | Call round | | | | |
|---|---|---|---|---|---|
| | 1 Verification | 2 Leadership team | 3 Teachers | 4 Families | All rounds |
| Calls assigned | 1,323 | 826 | 826 | 827 | 3,802 |
| Operators | 13 | 13 | 8 | 8 | 21 |
| Operator-days | 56 | 42 | 32 | 20 | 150 |
| Personnel cost, DOP | 73,141 | 54,777 | 41,870 | 28,544 | 198,331 |
| Personnel cost, USD | 1,386 | 1,038 | 794 | 541 | 3,759 |
| Cost per assigned call, USD | 1.05 | 1.26 | 0.96 | 0.65 | 0.99 |
| Cost per school assigned to a call, USD | | | | | 2.07 |

*Panel B. Cost per additional school completing each step, any reminder call*

| | Effect of any call, percentage points | Cost per additional school, USD | |
|---|---|---|---|
| | | Estimate | 95 percent range |
| Opened the email | 19.3 | 10.7 | 9.0 to 13.2 |
| Downloaded the report | 14.1 | 14.6 | 11.5 to 20.2 |
| Documented a workshop | 26.9 | 7.7 | 6.8 to 8.7 |
| Submitted an improvement plan | 8.1 | 25.5 | 20.0 to 35.2 |

*Panel C. By number of calls assigned*

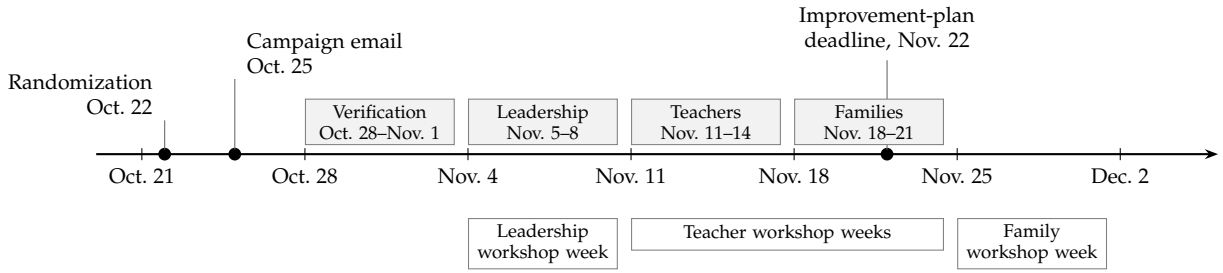
| | 1 call | 2 calls | 3 calls | 4 calls |
|---|---|---|---|---|
| Cost per school, USD | 0.99 | 1.98 | 2.97 | 3.95 |
| Effect on documenting a workshop, percentage points | 17.2 | 25.6 | 36.0 | 43.0 |
| Cost per additional workshop, USD | 5.8 | 7.7 | 8.2 | 9.2 |
| Effect on submitting a plan, percentage points | 4.7 | 8.9 | 9.7 | 14.7 |
| Cost per additional plan, USD | 21.1 | 22.2 | 30.7 | 26.9 |

*Notes:* Costs are the personnel cost of the call center, valued from the IDEICE payroll of November 2019, the month of the campaign. An operator is charged for every day on which the call logs record at least one call by that operator in that round, at the daily equivalent of the monthly salary (monthly salary times 12 over 260 working days), so a day with a single logged call is charged in full and the charge overstates the time spent on calls. The call-center coordinator is charged for every day with any logged call, including the contractor round. Operator-days and the cost per operator-day exclude the coordinator. Rounds 1 to 3 were placed by 13 IDEICE operators. One operator has no payroll row and is priced at the contract-operator salary. Round 4 was placed by 8 callers of an external contractor whose invoice is not available, and is priced at the average in-house cost per operator-day of rounds 1 to 3. Costs are salaries only. Adding employer social-security contributions of 15.4 percent (Law 87-01) raises every figure by that share, to 228,855 DOP in total. Not costed are telephone charges, the email campaign, MINERD and IDEICE staff time outside the call center, principals' time, and the assessments themselves. Dollar figures use the Banco Central average rate for November 2019, 52.76 DOP per USD. Panel B divides the cost per treated school (2.07 USD, the total over the 1,820 schools assigned to a call) by the intent-to-treat effect of any reminder call in Table 2, and the range divides by the ends of the effect's 95 percent confidence interval. Panel C prices a school by its assigned number of calls at the average cost per call, 0.99 USD, and divides by the corresponding effect in Table 3.

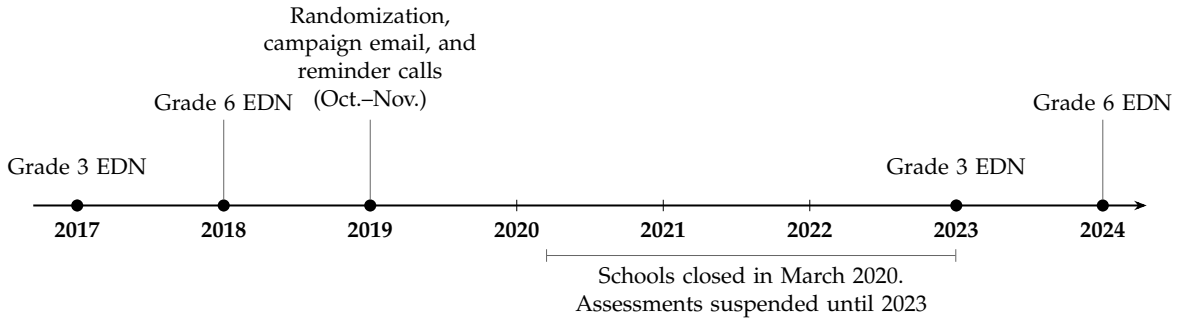
# Appendix C. Figures

Figure A1: Implementation Timeline

(a) The 2019 campaign, week by week

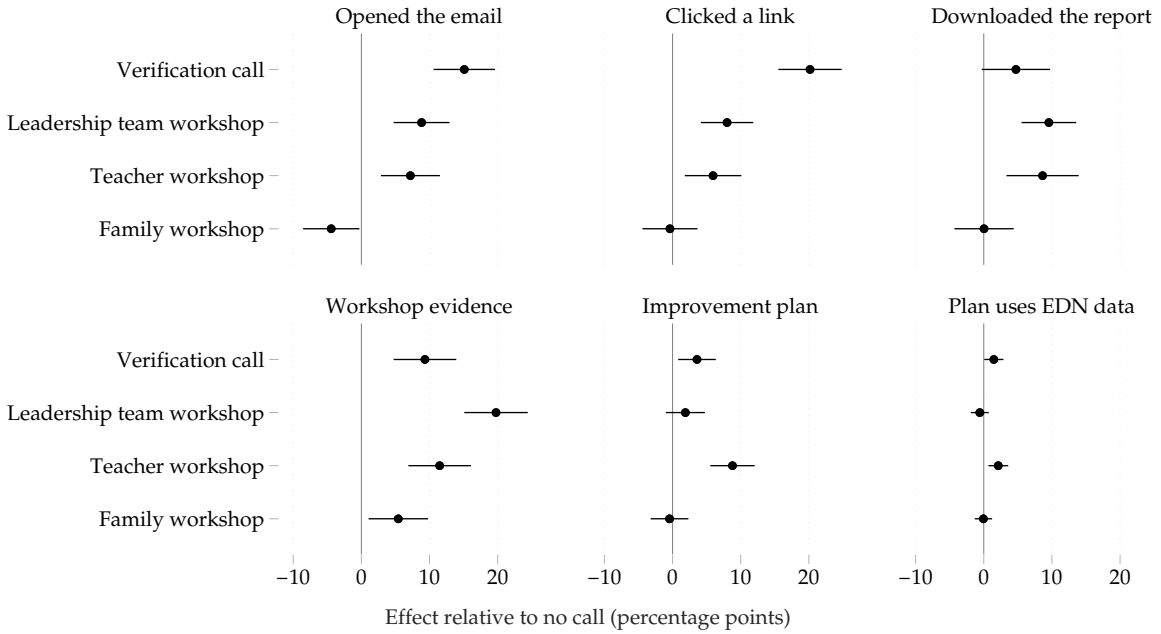



(b) Diagnostic assessments and the intervention, 2017 to 2024




*Notes:* Panel (a): shaded bands are the four reminder-call rounds, labeled by the workshop each one promoted, with the call dates recorded in the operator logs. Open bands are the workshop weeks that MINERD's email of October 2019 assigned to schools, and each reminder round preceded or coincided with its workshop week. Panel (b): each EDN is a cross-section of a different cohort. The grade 3 EDN of 2017 and the grade 6 EDN of 2018 are the baselines, and the grade 3 EDN of 2023 and the grade 6 EDN of 2024 are the outcomes. Schools closed in March 2020 and the national assessments were suspended until 2023.

Figure A2: Effects of Reminder Content, by Step of the Implementation Chain




*Notes:* Each marker is the effect of exposure to the indicated reminder call relative to the no-call control group, in percentage points, from the regressions reported in Table 4, in which the four call types enter jointly. Horizontal lines are 95 percent confidence intervals based on standard errors clustered at the district level. The count outcomes and the remaining measures of plan quality are reported in Table 4.