Call Me Educated: Evidence from a Mobile Phone Experiment in Niger

Jenny C. Aker and Christopher Ksoll*

September 2018

Abstract. What is the most cost-effective way to improve teacher accountability in remote areas of developing countries? We report the results of a randomized evaluation of an adult education program in Niger, in which a subset of villages were randomly assigned to a mobile phone intervention that called different stakeholders about teacher attendance. The adult education program was successful in improving students' reading and math skills by .19-.22 s.d., respectively. Students in villages with the additional mobile phone intervention achieved test scores that were .12-.15 s.d. higher as compared to those in adult education villages alone, with robust impacts for math across a variety of specifications. We present suggestive evidence on the mechanisms behind these results, distinguishing between the monitoring, motivational and reminder channels of the calls. The results suggest that the calls had a motivational effect on teachers and students, increasing teacher preparedness and decreasing students' likelihood of dropout. This suggests that using mobile phone technology as a means to communicate with teachers and students can improve learning outcomes, beyond its use as a pedagogical tool within the classroom.

JEL codes: D1, I2, O1, O3
Keywords: Adult education, teacher absenteeism, monitoring, information technology, Niger

*Jenny C. Aker, The Fletcher School and Department of Economics, Tufts University, 160 Packard Avenue, Medford, MA 02155; Jenny.Aker@tufts.edu. Christopher Ksoll, Mathematica Policy Research, Oakland, CA; christopher.ksoll@gmail.com. We thank Jonathon Robinson, James Berry, Michael Klein, Julie Schaffner, Shinsuke Tanaka and seminar participants at Tufts University, the Center for Global Development, IFPRI, the World Bank, University of Washington, Georgetown University, the University of Minnesota, the University of South Carolina. We are extremely grateful for funding from the DFID Economic and Social Research Council (Grant Number ES/L005433/1).

"Someone who works must be 'controlled'
Adult education teacher in Koutoukoutou, Niger

"(These calls) prove that our work is important."
Adult education teacher in Birji Zaoure, Niger

1. Introduction

In rural areas of developing countries, public sector absenteeism – of teachers, doctors, nurses or agricultural extension agents – is a widespread problem. In West Africa, teacher absenteeism in primary schools is estimated to be between 27-40% (Transparency International, 2013, Mbiti 2016). One potential solution has been to strengthen the monitoring of teachers (Banerjee and Duflo, 2006; Glewwe and Kremer, 2006, Mbiti 2016). Yet despite numerous interventions, such as community-based monitoring, hiring local teachers and audits, improving teacher accountability continues to be a significant challenge. This is particularly the case in countries with limited infrastructure and weak institutions, where the costs of monitoring are particularly high. For example, in Niger, the subject of this research, teacher absenteeism in a previous adult education program was 33% (Aker et al 2012).

The introduction of mobile phone technology throughout sub-Saharan Africa has the potential to reduce the costs associated with observing teachers' effort. By allowing governments and organizations to communicate with stakeholders in remote areas on a regular basis, digital technology may allow principals to more easily observe teachers' effort. Similarly, digital technology could allow communities to provide more timely feedback on teacher performance, thereby improving their engagement in the educational system.

To address these questions, we conducted a randomized evaluation of two interventions in rural Niger that were designed to improve teacher accountability and adult students' learning. Villages were randomly assigned to one of three groups. The first was a two-year adult education program (basic literacy and numeracy) in Hausa, with normal monitoring visits by non-governmental organization (NGO) and Ministry staff. The second included the same curriculum and monitoring as the first, but villages also received a mobile intervention: weekly phone calls to the teacher, village chief and two randomly selected students. A variant of this intervention was implemented in a subset of mobile villages during the second year, in which only teachers were called. The third was a pure control group, with no adult education program or calls. No other financial

incentives were provided for teachers in the short-term. The evaluation was carried out in 131 villages in the regions of Maradi and Zinder.

We find that both the adult education program and the additional mobile phone intervention improved students' learning outcomes. Across both years of the program, students in the standard adult education program increased their math and reading test scores by .19-.22 s.d., meaning that they were able to decode eight additional elements (letters, syllables, words) and solve two additional math problems. The mobile phone intervention led to an additional increase in math and reading test scores of .12 and .15 s.d., respectively, with stronger effects amongst called students. These effects are not driven by differential mobile phone ownership, nor changes in the composition of teachers over time. We also address alternative threats to identification, namely, differential attrition and baseline imbalance, and find that the math results are robust across these different specifications.

This evidence supports our hypothesis that, beyond its use as a pedagogical tool within the classroom, mobile technology can be used as an accountability device to improve students' learning. Yet a key question is the mechanisms through which these impacts occurred – as monitoring, motivation or reminders – and whether they primarily affected teachers or students. On the teacher side, teachers in villages that received calls were more likely to keep attendance logs and pass a competency exam. On the student side, students in the mobile intervention villages were less likely to drop out of the course, and called students had substantially higher test scores in the short- and medium-term. While the mobile intervention did not increase the extensive or intensive margin of teacher attendance, given the quality of our attendance data, we cannot rule out the possibility that the calls also affected teacher absenteeism.

Nevertheless, from a policy perspective, the magnitude of the effects on students' learning in mobile monitoring villages is relatively large, especially for called students, and cost-benefit calculations suggest that the calling intervention was cost effective. The question, however, is whether it is necessary to call all three stakeholders – the teacher, students and village chief – or just a subset. While we implemented a variant of the mobile intervention in the second year to test this hypothesis, we did not find differential effects between the two interventions. While this could be due to limited power, the learning effects suggest that simple calls could be a promising tool for policymakers to improve learning outcomes in rural areas.

Our results contribute to several strands of literature. First, despite decades of investment in adult education programs in developing countries, to our knowledge, it is one of two randomized evaluations of these programs (Banerj et al 2017). Aker et al (2012) evaluated the impact of a mobile phone-enhanced adult education program in Niger, finding that using mobile phones as pedagogical tools increased writing and math scores as compared to a standard adult education program. However, there was no pure control group. More recently, Banerji et al (2017) evaluated the impact of a maternal literacy and at-home learning program in India, finding that these interventions increased mothers' test scores by .05-.12 s.d.[1]

Second, our study contributes to the existing literature on the effectiveness of teacher monitoring, which primarily focuses on teachers of school-aged children (Guerrero et al 2013, Mbiti 2016). Duflo, Hanna and Ryan (2012) and Cueto et al (2008) find that monitoring programs combined with financial incentives reduce teacher absenteeism, with the former study also finding improvements in children's test scores. Muralidharan et al (2017) find that increased school monitoring is strongly correlated with lower teacher absence. Cilliers et al (2018) find that local monitoring combined with financial incentives improves teacher attendance.[2] Our experiment is somewhat unique in that it did not provide any explicit financial incentives for teachers, which may be easier for governments to implement, and focuses on an adult education program.

Finally, our paper contributes to the literature on community-based monitoring and inspection systems (Svensson 2007, Olken 2007, Bengtsson and Engstrom 2013). There is extensive evidence showing that material and financial incentives can crowd out intrinsic motivation if agents are mission-oriented rather than self-interested, thereby suggesting that incentives or monitoring can lead to a null or negative effect on agents' effort. At the same time, recent work suggests that extrinsic incentives may not crowd out intrinsic motivation (Bengtsson and Engstrom 2013). Our study suggests that the latter effect may be at work, as teachers reported that "these calls…show that our work is important".

[1]Royer et al (2004) evaluate the impact of different instructional approaches in the context of an adult education program in Burkina Faso.

[2]De Ree et al (2018) estimate the impact of an unconditional doubling of teachers' salaries in Indonesia, finding an improvement in teachers' job satisfaction but no impact on teacher effort or students' learning outcomes. Assessing the impact of a contract teacher program, Duflo, Dupas and Kremer (2015) find that test scores increased for students assigned to be taught by locally-hired contract teachers, potentially because of their low absence rates.

The remainder of the paper is organized as follows. Section II provides background on the setting of the research and the research design, whereas Section III presents the conceptual framework through which this intervention might affect student learning. Section IV describes the different datasets and estimation strategy, and Section V presents the results. Section VI provides evidence on the mechanisms and Section VII discusses alternative explanations. Section VIII discusses cost-benefit analyses and Section IX concludes.

2. Research Setting and Experimental Design

With a gross national income per capita of $641, Niger is one of the lowest-ranked countries on the UN's Human Development (UNDP 2016). The country has some of the lowest educational indicators in sub-Saharan Africa, with an estimated literacy rate of 15 percent in 2012 (World Bank 2015). Illiteracy is particularly striking among women and within our study region: It is estimated that only 10 percent of women attended any school in the targeted regions.

2.1. Program Description

2.1.1. *Adult Education Program*

Over a two-year period (2014 and 2015), an international NGO, Catholic Relief Services (CRS), and the Ministry of Non-Formal Education implemented an adult education program in two rural regions of Niger. The intervention was designed to provide five months of literacy and numeracy instruction per year to illiterate adults, with a total of 10 months of instruction over two years. Courses were held between February and June, with a break between July and January due to the agricultural planting and harvesting season. All classes taught basic literacy and numeracy skills in the native language of the village (Hausa), as well as functional topics on health, nutrition and agriculture. The curriculum was designed to introduce simpler reading and math tasks in the first year, and move on to more complicated tasks in the second year.[3]

[3]The same students were supposed to stay in the class for the two-year period; if a student dropped out in the first year, the teacher was not supposed to accept any new students. In practice, however, if an original student dropped out, the teacher may have accepted a new student into the class, thereby dividing the teacher's attention across different literacy and numeracy levels.

Each village was allocated 50 slots for the adult education program, with spots for 35 women and 15 men.[4] These fifty students were taught in two literacy classes, separated by gender. Since men's and women's classes differed by both gender and class size, we are unable to disentangle the potential differential effects of gender and class size on learning outcomes. Each class was held five days per week for three hours per day, and was taught by community teachers who were selected and trained in the adult education methodology by the Ministry. The teacher was either from the community or a neighboring community, and taught both classes in the village.

Teachers were offered a five month contract, with the possibility of renewal. The total monthly salary was 35.000 CFA per month (US$62), in-line with salaries paid by other governmental and non-governmental adult education programs.

2.1.2. Mobile Monitoring Intervention

Despite a long history of adult education programs in Niger, such programs have often been plagued by low enrollment, high dropout and rapid skills depreciation. In prior research on an NGO-administered adult education program, Aker et al (2012) found that teacher absenteeism was 33%, with relatively higher rates amongst teachers who were not based in the same community.

Due to the large number of villages, as well as limited monitoring resources, a mobile monitoring component was implemented in a subset of adult education villages. Data collection agents from a local survey firm, Sahel Consulting, made weekly phone calls to four individuals over a six-week period, calling the teacher, the village chief and two randomly selected students (one female and one male, with different students randomly selected each year). Teachers were not informed in advance that they would be called, nor were they informed that students and the village chief would be called. Despite the fact that the same students were called each week (within a given year), 77% of teachers were aware that students had been called. No phones were

[4]This differs from our previous study, in which the 50 slots were equally allocated between men and women (Aker et al 2012). The donor for this program wanted to increase women's access to the adult education program, and thereby allocated more slots to women in each village.

provided to either teachers or students as part of this intervention, and mobile phone ownership was not a necessary condition to be eligible for the calls.[5]

During the weekly phone calls, the field agents used a script to introduce themselves and asked a series of questions: the number of classes held and the number of hours for each class; the number of students who attended; and whether the respondent had any additional information to share. The phone calls were introduced two months after the start of the adult education program, and neither students, teachers, nor CRS field staff were informed which villages were assigned prior to the intervention. CRS and the Ministry conducted their normal (in-person) monitoring activities in all of the adult education villages over the two-year period, regardless of the village's monitoring status.[6] In an effort to understand which stakeholders were the most important, the intervention was modified slightly during the second year: in a subset of monitoring villages, only the teachers were called ("light" monitoring), as opposed to the teacher, village chief and students ("full" monitoring).

While general information on the results of the monitoring calls were shared with CRS (without specifying the village or the teacher), due to funding constraints, neither CRS nor the Ministry were able to conduct additional follow-up monitoring visits beyond what they had previously planned for the year. In fact, the number of in-person monitoring visits was low over the two-year period, with each teacher receiving less than one visit per year. As a result, teachers were not formally sanctioned for less than contracted effort during the adult education classes; rather, teachers only learned whether they would be retained for the second year six months' later. Approximately 20 percent of teachers were replaced between the first and second years, some of

[5]During the initial enrollment phase of the program, phone numbers for the students, village chiefs and teachers were obtained. If an individual did not own a mobile phone, we asked for the number of a friend or family member, called this individual, and then interviewed the correct respondent. For the first year, the same two students were called over the six-week period; during the second year, two different students were randomly selected to receive monitoring calls.
[6]In-person monitoring visits by CRS and the Ministry usually involved short visits to the village at any time to see if the materials were in place, if the teacher was present and how far he or she had progressed. These visits did not necessarily occur during class time, and did not require classroom observations. In addition, since student decisions regarding dropout were usually made within the first six weeks of classes (Aker et al 2012), the monitoring intervention would not have affected the likelihood of dropout in the first year.

whom were fired. We observe a positive (but not statistically significant) correlation between a teacher's likelihood of being replaced and the mobile intervention (Table 7).[7]

2.2. Experimental Design

In 2013, CRS identified over 500 intervention villages across two regions of Niger, Maradi and Zinder. Of these, we first stratified by geographic region and sub-region and randomly selected 134 villages to participate in the research.[8] Among these 134 villages, we stratified by regional and sub-regional administrative divisions before randomly assigning villages to the adult education program **(**114 villages) or a control group (20 villages).[9] Among the 114 adult education villages, villages were then assigned to either the "mobile monitoring" or "no mobile monitoring" intervention, with half of the mobile monitoring villages assigned to the "light" monitoring condition for the second year. Three villages dropped out of the program due to chieftancy disputes prior to the start of the program. As a result, the final sample in this paper is 131 villages, 20 in the pure control group and 111 in the adult education program.[10] Despite the fact that we lost three villages after random assignment, the groups appear relatively balanced, suggesting that village-level attrition was orthogonal to the experimental treatment assignment. A timeline of the implementation and data collection activities is provided in Figure 1, and a map showing the villages and their treatment assignment is included in Figure 2. A figure showing the number of villages in each treatment is in Figure 3.

[7]CRS' criteria for firing teachers was twofold: 1) teachers' absence during random visits; and 2) lack of preparation. In practice, no formal sanctions for less than contracted effort were applied during the adult education classes (for example, no one was fired, pay was not reduced, no follow-up visits occurred). These decisions were made prior to the start of classes the following year.

[8]The original design identified 140 villages, 20 in the control group and 120 in the adult education program. Due to logistical and resource constraints, and prior to the baseline, this was reduced to 134 villages. We stratified by region and sub-region and randomly dropped six villages from the adult education and mobile monitoring villages, respectively. Regressing a binary variable for whether the village was dropped on a vector of village-level administrative characteristics yields a joint F-statistic of .53.

[9]While we only have 20 villages in the control group, our power calculations were based upon previous research in Niger. Aker et al (2012) found that an adult education program increased writing and math test scores to first and second-grade level; this non-experimental treatment effect was used as the basis of the power calculations.

[10]Amongst the 131 adult education villages, 57 are in the mobile monitoring condition and 54 are in the no mobile monitoring condition. We only have village-level administrative data on the three dropped villages. In the second year, half of the villages in the monitoring condition were in the "full" monitoring treatment (29), and half were in the "light" monitoring treatment (28). We test for balance with and without these villages using village-level administrative data (Table 1A).

Within each village, and prior to the baseline, CRS identified eligible students in both the adult education and control villages, for a total of fifty students per village. While this was intended to be 35 women and 15 men per village, in some villages, only women were selected for the program. Individual-level eligibility was determined by two criteria: illiteracy (verified by an informal writing test) and willingness to participate in the adult education program. The same recruitment process was followed for all villages, including the control villages, as these villages were supposed to be phased into the adult education program in 2016. If there were more than 50 eligible students per village, a village-level lottery was used to randomly choose students. Thus, our sample across all villages consists of those individuals who were selected by CRS to participate in the adult education program, regardless of whether or not they actually participated.[11]

3. Conceptual Framework

There are a variety of mechanisms through which the full monitoring intervention could affect students' learning outcomes. We outline each one of these in turn. Figure 4 provides an overview of the predictions with respect to the key outcomes that we measure.

3.1. Teachers

The primary motivation of the mobile intervention was to reduce teacher shirking, similar to that observed in a previous adult education program. Yet the calls might affect teachers' behavior in a number of ways, either because of the threat of firing, reminders or motivation, or some complementarities between the three.

Shirking. In the classic principal-agent framework, Holmstrom (1979) shows that the optimal contract between a principal and agent will make use of information from additional signals on the agent's action, which in our case is obtained through phone calls. In our setting, the NGO and Ministry could have used the information obtained from the calls to fire teachers who were shirking, as the calls could have increased the probability of shirking being detected. They did not use this information, as we did not independently verify whether the teachers were present or conduct follow-up visits. Nevertheless, as long as teachers believed that

[11] While the random assignment to different treatment status occurred prior to student selection, neither CRS nor the survey teams were aware of the treatment status of each village.

they *could* be fired or penalized for absenteeism, the calls should increase effort and reduce shirking, if the employment was worth preserving.[12]

Teacher-level data suggest that the phone calls were understood as attempts to monitor teacher effort along the lines of the principal-agent model. Approximately 40% of teachers in monitoring villages stated that "Someone who works must be controlled", and felt that the calls had a monitoring purpose.[13] After the first year, 70% of teachers thought the calls were from CRS, whereas 29% thought that they were from the Ministry, suggesting that teachers thought that the calls were from supervisory figures. Teachers were also generally aware of the other calls: 80% of teachers reported knowing that the village chief was called, and 77% knew that some students were called.

Motivation. A second set of mechanisms through which the monitoring intervention could have affected teacher behavior is related to motivation. There is extensive theoretical and empirical evidence showing that material and financial incentives can crowd out intrinsic motivation if agents are mission-oriented rather than self-interested, thereby suggesting that incentives or monitoring can lead to no or a negative effect on agents' effort (Frey and Gee 1997, Frey and Jegen 2001). At the same time, recent work suggests that extrinsic incentives may not crowd out intrinsic motivation (Bengtsson and Engstrom 2013), and may increase motivation if the incentives are perceived as encouragement (Brock et al 2016). In our context, the latter scenario seems to be the case, as 60% of teachers in monitoring villages stated that "these calls…show that our work is important" and "encourage us to do well". If the calls increased teachers' perception of the importance of their work, this could, in turn, increase their motivation and effort, thereby decreasing absenteeism or increasing their willingness to prepare.

Reminders. A third mechanism through which the calls may have affected teachers is as a "reminder": the calls may have served as reminders about the courses, as they focused primarily focused on whether the classes had taken place and the number of hours held. Similar to the health literature, which has found that SMS reminders can increase medication adherence (Sarkar et al. 2015, Gurol-Urganci et al. 2013), the calls could

[12]As the calls only asked about whether a teacher held a class and the number of hours taught, in theory, any observed effect should be along those margins.
[13]Since the calls also involved village chiefs, this could have increased their oversight of the adult literacy classes.

have reminded teachers that they needed to hold classes. While this is a possible mechanism, the classes were held every day, similar to typical formal employment in Niger.

3.2. Students

As two students in each classroom were assigned to receive monitoring calls, the intervention could have affected students in similar ways. First, called students could have felt more accountable to attend class, thereby reducing the likelihood of dropout and increasing attendance. Second, the calls could have motivated students by making them feel as if their studies were important, thereby encouraging them to put in more effort to attend class or study more outside of class.[14] And finally, the calls could have served to remind students to attend class or arrive on time. Anecdotally, teachers in monitoring villages noted that students were more likely to attend and to show up on time for their courses, although we do not have similar data in non-monitoring villages.

While these effects should have a direct effect on called students, the calls could have had indirect effects on other students or the teacher, thereby raising learning outcomes for non-called students. This is akin to the spillovers associated with a girls' scholarship program in Kenya (Kremer, Miguel and Thornton 2009).

3.3. Teacher and Student Interactions

Because the mobile monitoring intervention called several stakeholders during the first year, it is likely that teachers' behavior of affected students, and vice versa. For example, more present and motivated teachers might, in turn, make students more excited to attend class and study more; similarly, more motivated students may have encouraged teachers to invest more in their preparation or hold more classes for longer.

4. Data and Estimation Strategy

The data we use in this paper come from four primary sources. First, we conducted individual student-level math and reading tests before, during and at the end of the program, and use these scores to measure the impact of the program on educational outcomes. Second, we implemented household-level surveys before and after the program. Third, we collected administrative and survey data on teachers, and use these data to better understand how the intervention affects teachers' effort. Fourth, we collected attendance logs from the adult

[14]Students thus may have increased their effort along both observable (participation) and unobservable (studying at home) dimensions (Jensen 2012, Attanasio and Kaufman 2014).

education centers in order to better understand if the intervention affected teacher attendance. Before presenting our estimation strategy, we discuss each of these data sources in detail.

4.1. Student and Household Data

4.1.1. Student Test Scores

As outlined above, CRS identified students in all villages and for all cohorts in January 2014. While we had originally intended to implement the baseline in all villages, the delayed start of the program during the first year, combined with delays in research funding, meant that we were only able to conduct the baseline in 91 villages.[15] In these villages, we stratified students by gender and took a random sample of 15 students per village, 10 women and 5 men. We implemented reading and math tests prior to the start of courses (February 2014), with an intended baseline sample of 1,365 students, excluding attrition. We administered follow-up tests in the baseline villages as well as the remaining villages in August 2014 and August 2015, thereby allowing us to estimate the immediate learning impacts of the program. This total *intended* sample was 1,965 students across 131 villages, excluding attrition.[16]

To test students' reading and math skills, we used USAID's Early Grade Reading Assessment (EGRA) and Early Grade Math Assessment (EGMA) tests. EGRA is a series of timed tests that measure basic foundational skills for literacy acquisition: recognizing letters, reading simple words and phrases and reading comprehension (Dubeck and Gove 2015). Each task ranges from 60-180 seconds; if the person misses four answers in a row, the exercise is stopped. EGMA measures basic foundational skills for math acquisition: number recognition, comparing quantities, word problems, addition, subtraction, multiplication and division (Reubens 2009).[17]

---

[15]To choose baseline villages, we stratified by region, sub-region and treatment status and selected a random sample. The number of baseline villages was equally proportioned between monitoring and non-monitoring villages (41 in each treatment). However, there were only 9 control villages, which means that some strata have no control villages in the baseline data. Thus, the mean and s.d. of the control group are estimated off of the learners in a small number of clusters that are not equally proportioned amongst strata.

[16]For the baseline, approximately 80 students could not be found and there are some missing observations. For the follow-up samples, in addition to attrition, there are approximately 60 missing values for the reading and math tests. This is in part due to refusals (between 6-10 people) and missing observations.

[17]The baseline EGMA test followed the standard procedure, which was to read the math questions aloud and allow for verbal responses. Yet we only included a subset of the simpler math tasks. For the follow-up math tests, we included the full battery of tests and asked respondents to solve the math problems on paper, in order to measure their ability to decode written information. Thus, the baseline and follow-up math tests are not directly comparable, which slightly affects our estimation when using the ANCOVA specification.

The EGRA and EGMA tests were our preferred survey instruments for two reasons. First, most adult education programs are criticized for high rates of skills' depreciation. Yet these high rates of skills' depreciation may be simply due to the lack of *automaticity* of reading achieved by the end of adult education programs, which are often not captured in traditional untimed tests. Thus, the EGRA timed tests allow us to determine whether participants in adult education classes are attaining the threshold required for sustained literacy acquisition. In addition, the tests allow us to measure the skills necessary for reading acquisition, such as simple decoding and reading comprehension (Dubeck and Gove 2015).

During the reading and math tests, we also measured students' self-esteem and self-efficacy, as measured by the Rosenberg Self-Esteem Scale (RSES) and the General Self-Efficacy Scale (GSES). The RSES is a series of statements designed to capture different aspects of self-esteem (Rosenberg 1965). Five of the statements are positively worded, and five are negatively-worded. Each answer is assigned a point value, with higher scores reflecting higher self-esteem. The GSES is a ten-item psychometric scale that is designed to assess whether the respondent believes that he or she is capable of performing new or difficult tasks and to deal with adversity in life (Schwarzer and Jerusalem 1995). The scale ranges in value from 12-60, with higher scores reflecting higher perceived self-efficacy. We use these results to measure the impact of the program on participants' perceptions of self-esteem and self-efficacy.

Survey attrition is a concern in most studies, especially in populations that engage in seasonal migration. Table A1 formally tests whether there is differential attrition by treatment status in 2014 and 2015, whereas Table 1 checks for balance amongst non-attriters in the baseline sample. The rate of attrition in the control group was 5 percent in the first year, with no statistically significant difference in attrition between the "any adult education" and control villages (Panel A). During the first year, there was relatively higher attrition in the adult education group relative to the control group, and lower attrition in the mobile monitoring group. The results are similar once controlling for other characteristics, such as gender (Panel C). This suggests that the

monitoring intervention might have discouraged survey attrition, at least in the first year. Women were less likely to attrit than men, with a marginally statistically significant difference by treatment status (not shown).[18]

The rate of attrition in the control villages was higher in the second year (8 percent), but there was no differential attrition by treatment status. While women were still less likely to attrit, this was not differential by the adult education or monitoring intervention (not shown). Overall, during the first year, the difference in attrition by gender may bias our treatment effect for the adult education program downwards, since female students had lower test scores as compared with male students (Aker et al 2012). Lower attrition in mobile monitoring villages could bias our treatment effect upwards or downwards, depending on whether individuals with higher or lower outcomes were more likely to attrit. To correct for potential selection bias due to non-differential attrition, we bound our treatment effects using Lee bounds and the inverse Mills' ratio (Table A7).

4.1.2. Household Surveys

The second dataset is household surveys conducted in February 2014 and December 2016, immediately prior to (and six months' after) the end of the program. The surveys collected detailed information on household demographics, assets, production and sales activities, access to price information, migration and mobile phone ownership and usage. During the final survey, we also conducted a simple timed syllable and word reading test. These data are primarily used to test for baseline imbalances across the different treatments, as well as impacts of the program on mobile phone ownership and learning six months after the end of the program.

4.2. Teacher Data

4.2.1. Administrative Data

The first teacher-level dataset is administrative data from CRS' recruitment and training procedures for the adult education program. This database includes information on teachers' education, age, gender, marital status, work experience and village of residence. These data allow us to not only conduct baseline balance checks, but also to determine whether the composition of teachers changed between the first and second years of

[18]The p-values testing for differences between women in the adult education and mobile monitoring interventions is .08 in the first year, and .79 in the second year. In the first year, women were less likely to attrit in the adult education treatment group as compared with women in the pure control group.

the program. In addition, we also have data on scores from a post-training exam, which all teachers were required to take before being formally hired.[19] We use the post-training exam scores from the beginning of the second year to measure whether the monitoring intervention had an impact on teachers' preparation before the start of the second year.

4.2.2. Teacher Survey Data

In addition to the administrative data, we conducted surveys with teachers in August 2014 and August 2015, at the same time as the student follow-up tests. These surveys included information on teachers' socio-demographic characteristics, self-reported absence and experience with the program.

Table A2 shows teacher survey attrition by year. Overall, 15% of teachers were not interviewed in 2014 in the adult education villages. While teachers from mobile monitoring villages were less likely to have attrited, this difference in not statistically significant at conventional levels (Panels A and B). In addition, teachers who had some secondary education were less likely to attrit during the first year. While the rate of attrition is similar in the second year (15%), there is no differential attrition by treatment status (Panels A and B).

4.2.3. Teacher Attendance Logs

Our final dataset is comprised of teachers' attendance logs, which were recorded by teachers and intended to be independently spot-checked and verified by CRS and the Ministry during unannounced visits. Unfortunately, these data were only collected by CRS during the second year of the program. In addition, these checks happened sporadically; as result, the data are potentially subject to recording bias, and so cannot be construed as an objective measure of their attendance. For example, if monitoring teachers correctly reported their attendance as compared with non-monitoring teachers, these attendance logs could underestimate the effects of the monitoring program on teacher attendance. Despite potential issues regarding the quality of these data, we use them as an alternative measure of the impact of the monitoring intervention on teachers' attendance and motivation.

4.3. Pre-Program Balance

[19]The teacher exam data are only available for one region (Mayahi) in 2014, prior to the start of the program, and for both regions in 2015 (prior to the start of second-year courses). The first-year test scores were based upon a numeric score, whereas the second-year test scores were a ranking of "fail/needs work", "passable" and "good". We recodified these classifications from 0 ("fail") to 1 ("pass/good").

Tables 1A-1C present the baseline characteristics at the village, household and student level by treatment status, using data from the entire sample and amongst non-attriters. In addition, as the program was implemented at the village level, the standard errors are adjusted for clustering at the village level for those tables that have individual-level observations (Tables 1B-1C).[20]

Table 1A shows the estimated differences in means for a number of baseline village-level characteristics, using administrative data. The table is divided between the full initial sample of 134 villages, as well as the final sample of 131 villages included in the study. Average village size is 1200, and 70% of the population is from the Hausa ethnic group. Villages are located 6 km from a weekly market, and only 10% of villages have access to an "improved" road. Approximately 50% of villages have a school, and only 5% have an adult education center. 80% of control villages had mobile phone access. Turning to the differences across groups, the groups appear well-balanced, with no statistically significant differences for these characteristics, with the exception of mobile phone access; adult education villages were more likely to have mobile phone access than both control villages and the mobile monitoring villages. The comparability is similar for the original sample of 134 villages and the restricted sample of 131.

Table 1B shows the pre-program comparison of a number of student and household-level characteristics across 91 villages. The table is divided between the full baseline sample, as well as those in the baseline sample who also appear in either of the follow-up samples. Average age in the control group was 35 years. The average education level of household members was 1.79 years. Fifty-seven percent of households in the control sample owned a mobile phone at baseline, with 52 percent of control respondents having used a mobile phone in the months prior to the baseline. Differences in pre-program household characteristics by treatment status are small. While some baseline differences are statistically significant at the five percent level – such as asset ownership, which are related measures – the groups appear well-balanced overall.

Table 1C compares baseline reading and math z-scores by treatment status. Baseline reading z-scores are .07 s.d. in the adult education group and .19 in the mobile monitoring group, with slightly lower means when restricting the sample to non-attriters. For math, baseline z-scores are .15 s.d. higher in both the adult education

[20]We also conducted a test of the equality of means between villages in the "full" and "light" monitoring interventions in 2015, and did not find any statistically significant differences for any of the variables in Tables 1A-1D.

and mobile monitoring groups as compared to the control, with slightly smaller means when restricting the sample. The only statistically significant difference is between baseline reading z-scores in the monitoring group as compared with the control group, which is statistically significant at the 10 percent level. These differences could, in part, be explained by the small sample size in the control group. Nevertheless, the magnitude of some of these differences suggests that we should control for baseline test scores as a robustness check.

Table 1D presents a comparison of teacher characteristics by year, using teacher administrative data. Overall, teacher characteristics are well-balanced between the mobile monitoring and non-mobile monitoring villages. Teachers in adult education classes without mobile monitoring were 35 years old and 40 percent female; 44 percent had some secondary education. A majority of teachers were married and had a mobile phone prior to the start of the program, and approximately 60% were located in the same village as the adult education program. Teacher "exit" exam scores were similar between mobile and non-mobile monitoring villages prior to the start of courses in the first year, although this is based off of data from one region.

As approximately 20 percent of teachers were replaced between the first and second year – either by their own volition or because they were fired – it is important to test for differences in teacher characteristics over time. For example, if the monitoring program weeded out "poorer" teachers, then this could have changed the types of teachers in monitoring and non-monitoring villages during the second year. While there are some changes in the composition of teachers over time – notably, there were fewer "local" teachers and almost every teacher had a mobile phone – teachers' characteristics were similar between treatments (Table 1D, Panel B). Overall, these results suggest that while the composition of teachers changed over time, it did not change differentially by treatment status.

4.4. Estimation Strategy

To estimate the impact of both the adult education and mobile phone interventions on educational outcomes, we use a simple differences specification. Let $test_{iv}$ be the reading or math test z-score attained by

student *i* in village *v* immediately after the program.[21] $adulted_v$ is an indicator variable for whether the village *v* was assigned to the adult education intervention (*adulted*=1) or the control (*adulted*=0). $adulted*monitor_v$ takes on the value of one if a village was assigned to the adult education program augmented with any mobile phone intervention, and 0 otherwise. θ_S are geographic fixed effects at the regional and sub-regional levels (the level of stratification). We include a year fixed effect and pool observations across the two years to estimate the following specification:

$$(1) \qquad test_{ivt} = \beta_0 + \beta_1 adulted_v + \beta_2 adulted_v * monitor_v + \theta_S + \theta_{2015} + \varepsilon_{ivt}$$

The coefficients of interest are β_1 and β_2, which capture the average immediate impact of being assigned to the adult education program (without phone calls) as compared with the control, as well as the additional impact of the mobile phone intervention, respectively. We cluster the error term at the village level for all student-level specifications. However, given the small number of clusters in the pure control group, we also bootstrap our standard errors as a robustness check.[22]

While equation (1) is our preferred specification, we also estimate the impact of the program using a value-added specification. However, the value-added specification reduces our sample size considerably, as we only have baseline data for 91 villages. Thus, in addition to the standard value-added specification, we also impute a zero value for missing baseline observations and include a variable for baseline availability.

In addition to the above equation, we also estimate two additional specifications to account for the design features of the program. First, when estimating the impact of the program in 2015, we include binary variable for both the "full" and "light" monitoring interventions. Second, since a subset of students were called in each mobile monitoring village in the first year, and in the "full" monitoring villages during the second year, we estimate the following regression:

[21]There are a number of ways that raw EGRA and EGMA scores can be used and transformed for analysis, including raw untimed scores, raw timed scores, untimed normalized scores and timed normalized scores. The results in the tables show the timed normalized scores for reading and the untimed normalized scores for math. Results are robust to using raw non-normalized scores (as provided in the Appendix) and untimed normalized scores for reading. All test scores are normalized to the contemporaneous control distribution, following Aker et al (2012). They are also robust to normalizing test scores by region.

[22]Results are robust to including a binary variable for the ABC program, an additional program that was supposed to be implemented in 2016, after the end of this program. Regressions do not include a binary variable for female, as a number of villages only included female students.

2) $$test_{ivt} = \beta_0 + \beta_1 adulted_v + \beta_2 adulted_v * monitor_v + \beta_3 adulted_v * monitor_v * called_{iv} + \theta_S + \theta_{2015} + \varepsilon_{ivt}$$

where $adulted_v$*$monitor_v$*$called_{iv}$ is equal to 1 if student *i* in village *v* was assigned to receive a weekly call in a mobile village. In this specification, β_1 is interpreted as in equation 1, whereas β_2 is the average additional impact of the mobile intervention on all non-called students in the classroom. β_3 is the average additional impact of being called, which is in addition to the impact of the mobile intervention.

5. Results

Figures 5A and 5B depict the mean normalized reading and math test scores by treatment status. Test scores are normalized using the mean and s.d. of contemporaneous test scores in control villages. Three things are worth noting. First, the adult education program increases reading and math scores significantly as compared to the control group, with relatively stronger effects on reading. Second, the impacts of the adult education program are stronger for simpler reading tasks, namely, letter or syllable recognition. For math, however, the impacts of the adult education program are stronger for more difficult math tasks, such as addition, subtraction, multiplication and division, as some students were able to recognize numbers prior to the program. And third, the difference in test scores between mobile monitoring and non-mobile monitoring villages are substantial in magnitude, especially for tasks.

5.1. Pooled Impacts of the Intervention

Table 2 presents the results of Equation (1) for reading (Panel A) and math z-scores (Panel B) across both years of the program, focusing on the composite scores. For reading, the adult education intervention increased students' reading test scores by .22 s.d. over the two-year period, with statistically significant effects at the 5 percent level (Table 2, Panel A). By the end of the program, students in the standard adult education classes could successfully complete eight reading tasks (reading letters, syllables and words) as compared to those in the control villages (Table A3, Panel A). Nevertheless, the program was not successful in raising students' reading scores to threshold reading level of 1 word per 1.5 seconds, an indicator for sustained learning.

The coefficient estimate of "*adulted*monitor*" variable is the additional effect of being in a village assigned to the mobile intervention. For reading, the mobile intervention increased students' test scores by an additional .15 s.d., with a statistically significant effect at the 10 percent level. By the end of the program,

students in the mobile villages were able to complete an additional six reading tasks as compared to those in the standard adult education villages, with a statistically significant impact at the 10 percent level (Table A3, Panel A).

The results are similar for math z-scores (Table 2, Panel B): the adult education program increased math z-scores by .19 s.d. as compared with the control group, with statistically significant effects at the 5 percent level. Concretely, this means that students were able to correctly complete two additional math problems by the end of the program (Table A3, Panel B). The mobile intervention further increased students' math scores by .12 s.d., with statistically significant effects at the 10 percent level (Table 2, Panel B). Concretely, this implies that students in the mobile monitoring villages were able to complete an additional two math tasks as compared to those in the standard adult education villages by the end of the program (Table A3, Panel B).

As there was some imbalance in the baseline reading and math z-scores, we may be concerned that our results are being driven by baseline differences, especially for reading. Yet the baseline was only implemented in 91 villages, thereby reducing our power. We address this in two ways. First, we estimate the standard ANCOVA specification, using observations from only the 91 baseline villages (Table A4, Column 1). Second, we estimate an ANCOVA by imputing a zero baseline value for those students in non-baseline villages (Table A4, Column 2) and including a binary variable indicating whether the baseline test score was missing. Overall, the impacts of the mobile intervention are robust for reading and math z-scores when using the full sample and the imputed zero value (Table A4, Column 2). The point estimate on the mobile intervention is smaller (and not statistically significant) for reading when restricting the sample to the baseline villages, but remains positive and statistically significant for math (Table A4, Column 1). This is unsurprising, as there was no baseline difference in math scores between the adult education and mobile treatments. Thus, while the results persist for math after controlling for baseline imbalance in the restricted sample, we are unable to reject whether the null results for reading are due to baseline imbalance or the small sample size.[23]

5.2. Effects of the Program over Time

[23] As baseline data were not collected for the more complex math tasks (difficult addition and subtraction and multiplication and division), we control for the baseline value of the simpler addition/subtraction subtask in Columns 4-5 of Table A4 (Panel B).

While the results in Table 2 suggest that the mobile monitoring intervention increased reading and math z-scores as compared with the standard adult education program, a key question is the dynamics of these effects over time, once teachers learned about the monitoring intervention and adults achieved higher learning outcomes. Tables 3 and 4 show the impacts of the adult education and mobile monitoring interventions by year for reading and math, respectively.[24]

The standard adult education program increased students' reading z-scores by .22 s.d. during the first year, with a statistically significant effect at the 5 percent level (Table 3, Panel A). The mobile monitoring intervention increased students' reading z-scores by an additional .19 s.d., with statistically significant effect at the 5 percent level (Table 3, Panel A). Panel B shows the results for 2015, breaking out the monitoring intervention between "full" and "light" treatments. Reading z-scores in the standard adult education program are of similar magnitude during the second year (.22 s.d.). While the coefficients on the "full" and "light" monitoring interventions are positive, they are not individually or jointly statistically significant, and they are not statistically significantly different from each other. In addition, we do not find a statistically significant difference between the coefficients on the monitoring intervention for the first and second year (p-value of .13).

Turning to math, the adult education program had positive and statistically significant impacts on students' math scores (Table 4): In the first year, the standard adult education program increased students' test scores by .15 s.d., increasing to .23 s.d. in the second year, with a statistically significant difference between the two (p-value of .07). This suggests that students improved their math skills over time in the standard adult education program. Similar to reading, the monitoring intervention had positive and statistically significant impacts in the first year (Table 4, Panel A), increasing math scores by an additional .14 s.d., with a statistically significant effect at the 10 percent level. In 2015 (Panel B), the monitoring calls did not have individual or joint statistically significant effects on math z-scores. We also do not find a statistically significant difference between the coefficients on the monitoring intervention for the first and second year (p-value of .40).

[24]In addition to estimating separate regressions by year, we also estimate a regression that includes the interaction between *adulted*time* and *adulted*monitoring*time*, using data from both years. The p-values are reported in the text.

Overall, the results in Tables 3 and 4 suggest that the adult education program was successful in improving students' learning outcomes over time for math, although not necessarily for reading. While the monitoring intervention led to additional improvements in test scores in the first year, there were no additional benefits to the monitoring intervention in the second year. Why might this be the case? First, if teachers or students in villages that received the calls learned that there were no sanctions or benefits after the first year, they could have reduced their effort, thereby making the monitoring intervention less effective over time.

A second potential explanation for similar impacts in the second year is that the "full" intervention – ie, calling teachers, students and the village chief – was necessary in order to lead to improvements in learning, perhaps due to increased student effort. Since only half of the original monitoring villages received the full monitoring intervention during the second year, we may be underpowered to detect any effects.[25]

5.3. Heterogeneous Effects by Students' and Teachers' Characteristics

We might expect greater impacts of the monitoring intervention according to students' and teachers' characteristics.[26] Table 5 tests for heterogeneous impacts of the program by whether the student was called, while Table 6 tests for heterogeneous effects by teacher characteristics.

5.3.1. Student Characteristics

Table 5 shows the results of the estimation of equation (2) across different years. In this specification, the coefficient on *adulted*monitor* represents the additional effect of the monitoring intervention on non-called students, whereas the coefficient on *adulted*monitor*called* represents the additional effect of the program on called students in monitoring villages.

Overall, Table 5 shows that the effects for non-called students are positive for both reading and math, but are only statistically significant for reading during the first year. In contrast, the effect of the mobile

[25]A way to test whether the similar effect of the monitoring intervention in the second year is due to teacher learning would be to identify teachers who shirked but were not fired. In theory, these teachers should exert less effort in the second year, as they were not sanctioned. However, not all teachers who left were fired, and we do not have shirking information on non-monitoring teachers.

[26]While the two regions involved in this study have similar agro-pastoral characteristics the Zinder region is closer to Nigeria and had a greater number of intervention villages. While the number of villages per field agent was similar between regions, the monitoring program could be more effective in improving learning outcomes if it was more difficult to travel to villages in Zinder. Table A5 reports the results of a regression that tests for differential effects of the monitoring program by region. The triple interaction term is not statistically significant for reading or math z-scores, suggesting that the monitoring program did not have differential impacts by region.

monitoring intervention on called students is 2-4 times as large as that for non-called students in monitoring villages, with statistically significant effects at the 5 and 10 percent levels for the pooled results (Panel A) and for the first year (Panel B).[27] Consistent with the results in Table 5, the impact of the monitoring intervention is slightly smaller once called students are excluded (Table A6), although the results are still positive and statistically significant for reading and math z-scores for the first year.[28] These results suggest that one of the mechanisms through which the monitoring intervention could have affected learning outcomes was via called students, who could have motivated other students or the teacher.

5.3.2. Teacher Characteristics

In a standard shirking model, a monitoring intervention should potentially have a stronger impact upon teachers who do not have an outside option. Table 6 presents the impact of the mobile intervention on reading and math z-scores by teachers' characteristics that proxy outside options, such as gender, level of education, and whether the teacher lived in the village. As we did not stratify our sample by these characteristics, we recognize that these results may be subject to substantial biases. Nevertheless, these estimates provide correlations between teacher characteristics and the mobile monitoring intervention.[29]

In many villages in our sample, women rarely migrate; as a result, female adult education teachers might have fewer outside options, thereby making the monitoring component more salient. This is confirmed by the teacher survey: While 46 percent of male teachers reported that they could find other work if they were not adult education teachers, only 24 percent of female teachers did so, despite the fact that both men and women had similar levels of education.

Overall, the gender of the teacher is not associated with improvements in learning outcomes. The mobile monitoring program is associated with positive improvements in learning for male teachers (Columns 1

[27]The specification for 2015 (Panel C) includes both full and light monitoring villages, although only students in full monitoring villages were called. If the variable "*monitor*called*" is modified to include students called in both 2014 or 2015, the coefficient is positive and statistically significant for reading and math, suggesting that the learning impacts on called students in the first year persisted into the second year.

[28]As two students were called in each of the monitoring villages in 2014 and in the full monitoring villages in 2015, there could be a maximum of 58-114 called students for each year. Our sample did not necessarily include all called students, however, and so the number dropped from this regression is less than the maximum number of called students.

[29]The number of observations is slightly less than the expected number of observations due to missing values for some teacher characteristics.

and 4), increasing reading and math z-scores by .11-.13 s.d. as compared with male teachers in the standard adult education villages. Students in monitoring villages with a female teacher had test scores that were .13-.17 s.d. higher, although not statistically significant. This suggests that the monitoring intervention was slightly more effective for female teachers.

In theory, teachers with higher levels of education should have better outside options, thereby reducing the effectiveness of monitoring component. In practice, the teachers in our sample have either primary or some secondary education, with very few teachers going beyond secondary school. Overall, students in classrooms with more highly educated teachers had higher test scores, increasing reading and math z-scores by .11-.13 s.d. respectively. The impact of the monitoring intervention on teachers with no secondary education was significant, increasing reading and math z-scores by .22-.23 s.d., with a statistically significant effect at the 5 percent level (Columns 2 and 5). The monitoring calls seemed to have no impact on more educated teachers.

While the mobile intervention could potentially make it easier for the community to observe teachers' absence, the nature of the intervention may have been more effective for local teachers, as they are subject to immediate social pressures within the community.[30] Students taught by local teachers had reading and math z-scores that were .22-.26 s.d. higher as compared to students with non-local teachers, with a statistically significant effect at the 1 percent level. The monitoring intervention played a similar role for non-local teachers, increasing test scores by .15-.22 s.d., with statistically significant effects. While there were no differential effects of the monitoring intervention by teacher residence, these results suggest that the monitoring intervention induced pressure on those teachers who were non-local, and hence potentially more likely to shirk.

6. Mechanisms

To understand the mechanisms behind these results, we estimate the impact of the mobile intervention on a variety of teacher and student measures. Table 7 presents evidence of the impact of the monitoring intervention on teacher attendance, as well as proxy measures for performance and motivation. Table 8 presents evidence on student dropout and motivation.

6.1. Teacher Behavior

[30]As shown in Table 1D, 64% of teachers in our sample were local in 2014, without a statistically significant difference by monitoring status.

As outlined above, the mobile phone intervention could have affected teachers' behavior via shirking, motivation and reminders. In order to test these mechanisms, we would ideally have high-frequency data on teacher attendance (including the duration of classes held), as well as teaching quality. However, due to limited in-person monitoring by CRS and the nature of the mobile intervention, we only have weekly attendance data from mobile monitoring villages. In addition, we do not have systematic data on teaching *quality* via classroom observations.[31] Therefore, we assess the impact of the monitoring intervention on teacher effort and motivation using a number of proxies. These data primarily come from teacher surveys, CRS administrative data and teachers' attendance logs.

Table 7 shows the results of the monitoring component on these indicators, presenting pooled results across both years (when available). Overall, 54% of teachers in adult education villages reported stopping the course at some point (Column 1). Yet the monitoring intervention did not have an impact on the extensive or intensive margin of stopping the course (Columns 1 and 2). While the monitoring intervention affected the *intensive* margin of shirking - teachers in monitoring villages reported stopping the course for .57 fewer days - this is not statistically significant (Column 2).[32]

Columns (3)-(5) capture alternative measures of teacher effort using teachers' attendance logs. Overall, two indicators are used: whether a teacher kept an attendance log (as attendance logs were not mandatory), and the number of classes recorded in the log. As the attendance log data were only collected for 2015, we are unable to test the impact of the intervention for both years. Overall, the monitoring intervention appeared to have an impact on the likelihood of a teacher keeping an attendance log: Teachers in mobile monitoring villages were 16 percentage points more likely to keep an attendance log, with a statistically significant effect at the 10 percent level (Column 3). While teachers were supposed to teach, on average, 20 days per month, the average number of days' taught in adult education villages was 9-11 days (Columns 4-5). Yet there was no impact of the

[31]A potential critique of the mobile monitoring intervention is that the observed changes are simply due to the Hawthorne effect; in other words, teachers are changing their behavior simply because they are being monitored. This is the precise purpose of the intervention and, we would argue, something that is inherent in all monitoring interventions, in-person or virtual.

[32]In mobile monitoring villages, there was a high intra-village correlation of responses amongst teachers, village chiefs and students, even when the teacher was absent; this did not appear to change over time. While this could be due to either collusion or a high degree of shared information amongst the stakeholders, we believe that this is less likely to be due to collusion, as students were also called.

monitoring intervention on the number of days taught, either pre- or post-calls. This could be due to the small number of villages that had attendance logs, or the fact that the attendance logs were only checked and verified sporadically. However, we are unable to determine whether the null effect is due to the absence of an effect, limited power or measurement error.

Column (6) assesses the impact of the program on teacher performance, namely, whether a teacher was replaced between the first and second year of the program. While 19 percent of adult education teachers were replaced, either due to firing or their own choice, and teachers in monitoring villages were 12 percentage points more likely to be replaced, there was no statistically significant difference between monitoring and non-monitoring villages. This is, perhaps, unsurprising, as CRS did not have access to the call data, and conducted few in-person monitoring visits to assess teachers' presence in the classroom.

A final measure is related to teachers' preparation within the classroom. Ideally, this would be measured by classroom observation data. As these data were not available, teachers were rated by the Ministry in January 2015, after the first year of courses and in preparation for the second year. The rating scale included "poor", "passing" and "good". Converting this to a binary variable ("0" for "poor" and "1" for "passing/good"), 83% of teachers in adult education villages passed the exam and were able to teach the second year (Column 7). Teachers in monitoring villages were 13 percentage points more likely to pass the exam, with a statistically significant effect at the 5% level. The coefficient is reduced to 11 percentage points when restricting the sample to the subset of teachers who were not replaced between the first and second year and is statistically significant at the 10 percent level.

Taken together, the results in Table 7 provide suggestive evidence that the monitoring intervention did not affect teachers' shirking, as there was no impact on teachers' attendance along the intensive or extensive margin. Rather, the monitoring program appeared to affect teachers' classroom preparation, as proxied by the likelihood of keeping an attendance log and their scores on the exit exam.

6.2. Student Behavior

While we do not have reliable high-frequency data on student attendance, we measure the impact of the monitoring intervention on student effort and motivation by using a number of proxies (Table 8). Overall, 28% of students dropped out of the course, with slightly higher rates of dropout in the second year. Across both years,

the monitoring intervention decreased the likelihood of dropout (Column 1), with a statistically significant impact at the 5 percent level (see the p-value at the bottom). A majority of those who dropped out reported doing so for reasons outside of their control, namely, pregnancy, illness or a death in the family. Nevertheless, the monitoring intervention reduced the likelihood of student dropout for an endogenous reason, primarily for called students, although these effects are not individually or jointly statistically significant.

The monitoring intervention could have affected students' self-esteem and self-efficacy, and hence their motivation for learning (Columns 3-4). Overall, adults in pure control villages had fairly high self-esteem and self-efficacy scores (as compared with the maxima of 20 and 40, respectively). Students in adult education villages had lower self-esteem and self-efficacy scores as compared with the control across both years, with negative and statistically significant effects on self-efficacy (Column 4). This is consistent with other literature, which finds that self-esteem and self-efficacy can be dynamic over time (Ksoll et al 2017). While the monitoring intervention appears to mitigate this negative effect for non-called students, none of the individual or joint effects are statistically significant. Thus, we cannot draw any conclusions about the effect of monitoring on these measures.

As a final measure, we assess the impact of the monitoring intervention on students' learning outcomes six months after the end of the program. This score was based upon a modified version of EGRA, whereby students were given a short time period (20 seconds) to read five words. After the program, the impact of the adult education program on reading is positive but not statistically significantly different from the control. This finding that is consistent with the literature on the rapid loss of skills in adult education courses, especially if adults do not achieve a minimum reading threshold – as was the case in our context. While there was no longer-term impact of the monitoring program on non-called students, there was a strong and statistically significant impact of the monitoring program on called students (Column 5). In fact, the coefficient on called students is four times larger than the point estimate for the adult education program alone.

Taken together, the results in Table 8 suggest that the monitoring intervention increased students' effort, as it decreased the likelihood of student dropout. These effects are primarily driven by the called students, leaving open the possibility that any observed effects on non-called students may be due to positive spillovers from called students, rather than from the teacher.

7. Alternative Explanations

There are several potential confounds to interpreting the above findings. First, there might be differential in-person monitoring between villages. If the Ministry or CRS decided to focus more of their efforts on mobile monitoring villages because they had better information, then any differences we observe in test scores might be due to differences in program implementation, rather than the mobile monitoring component. Yet during the implementation of program, there was very little in-person monitoring, and no differential visits by treatment status.

A second potential confounding factor could be due to differential attrition. The results on attrition in Table A1 suggest that attrition is higher in the adult education villages (as compared with the control group) and lower in the monitoring villages in the first year, although attrition is not differential during the second year. If students with lower test scores are the marginal survey attriters, then this could underestimate the impacts of the adult education program and overestimate the impact of the monitoring intervention.

As we are primarily concerned with this latter comparison, we use tightened Lee bounds and the inverse Mills' ratio to correct for potential bias due to differential attrition, restricting the sample to the adult education villages. We use both corrections, since Lee bounds are overly conservative. Table A7 presents these results. Unsurprisingly, the upper bounds are positive and statistically significant for both reading and math for all time periods. While the lower bounds are positive, they are not statistically significant (Columns 1 and 2). This is perhaps unsurprising, as most of the impacts of the monitoring intervention were only statistically significant at the 5 or 10 percent levels. Using the inverse Mills' ratio, however, the impact of the monitoring intervention for reading and math is positive and statistically significant across both years and the first year, although not the second year, similar to the full results.

Third, for some of the student and teacher survey measures, there could be concerns about non-classical measurement error, as teachers and students could systematically misreport. While this is an obvious concern for the survey data, which are used to measure dropout and attendance, this would be less of a concern for the test score data, as these are timed tests that objectively measure students' learning, and cannot be easily manipulated. In addition, while the teacher competency exam score data may be subjective across evaluators, this should not be correlated with monitoring status. Thus, while we are more concerned about the issue of non-

classical measurement error for the self-reported dropout and attendance measures, we are less concerned that this could be driving our results associated with student learning or teacher motivation.

Fourth, it is possible that higher baseline mobile phone ownership in monitoring villages could be driving differences in learning outcomes, or that the monitoring calls encouraged students or teachers to own more mobile phones, thereby increasing the relevance of their curriculum and improving learning outcomes (Aker et al 2012). While possible, data on mobile phone ownership over time and across treatments does not suggest this is the case. While average mobile phone ownership was higher in monitoring villages in the baseline, ownership and usage was similar across all treatments at endline (Tables 1B and 8). In addition, controlling for baseline mobile phone ownership does not affect the results, nor does controlling for changes in mobile phone ownership over time.

Finally, there might be differences in observable and unobservable characteristics in teacher quality. If the Ministry or CRS chose better-quality teachers for mobile monitoring villages, or better-quality teachers self-selected into those villages – which is unlikely, as this selection process happened before the randomization occurred - then any differences we observe in test scores might be due to differences in teachers' quality, rather than the presence of mobile monitoring program. The means comparison of teacher characteristics between treatments for each year of the program suggests that differences in teacher quality are unlikely to explain the results (Table 1D).

8. Cost-Effectiveness

A key question is the cost-effectiveness of the mobile intervention as compared to the standard adult education program. A previous evaluation of a similar adult education program in Niger cost around US$20 *per student* (Aker et al. 2010).[33] The mobile monitoring intervention cost (an additional) $3.08 *per village*, including the costs of agents' time and mobile phone credit, with a cost of $US.06 per student. As compared with learning gains in the standard adult education program, the mobile intervention led to an additional .12 s.d. in learning, primarily for math, across both years of the program, as compared with a cost of US$.06 per student.

[33] While in-person monitoring visits were limited over the duration of the study, we have data on per-monitoring costs for both in-person and mobile monitoring. On average, in-person monitoring costs are $6.20 per village, primarily including costs for the agent's time and gas for the motorcycle.

This is an order of magnitude more effective than most education interventions in developing countries (Evans and Ghosh 2008). Despite the fact that we are unable to precisely disentangle the mechanisms related to the learning outcomes, at this low cost per student, it suggests adult education programs should experiment with different digital formats to monitor and/or encourage teachers and students in remote rural areas.[34]

9. Conclusion

Adult education programs are an important part of the educational system in many developing countries. Yet the successes of these initiatives have been mixed, in part due to the appropriateness of the curriculum, the opportunity costs of adults' time and the ability of governments and international organizations to monitor teachers' effort.

This paper assesses the impact of an intervention that called teachers, students and the village chief as part of an adult education intervention in Niger. In addition to learning gains achieved during the adult education program, we find that these calls increased students' skills acquisition over the two-year period, especially for math and for called students.

While we provide insights into the mechanisms that explain these results, such as increased teacher motivation, lower student dropout and improved learning – especially amongst called students - we are somewhat constrained by our data on teacher attendance, which is the primary channel through which we would expect teachers to react if they are worried about being fired. Nevertheless, while disentangling the mechanism is difficult, given the cost effective nature of the intervention, this could potentially be a low-cost way to improve learning outcomes. In addition, it seems natural to experiment with a few different variants to better understand the mechanisms, such as linking the monitoring explicitly to firing or pay decisions (along the lines of (Cilliers et al. 2018); trying to increase the motivational effects of the calls by being more explicitly supportive (see Brock et al. 2016); or varying the number of called students in a classroom.

[34] Interestingly, the tradeoff is somewhat less clear for the group for which the benefits are strongest and most consistently statistically: per called student, the cost was about US$1.5 for an additional .28-.37 s.d. in additional learning. Calling every student in the class would add significantly to the cost of the program and might not be cost-effective.

References

Abadzi, Helen. 1994. "What We Know About Acquisition of Adult Literacy: Is There Hope?," In *World Bank discussion papers,*, ix, 93 p. Washington, D.C.: The World Bank.

Abadzi, Helen. 2013. Literacy for All in 100 Days? A research-based strategy for fast progress in low-income countries," *GPE Working Paper Series on Learning No. 7*

Aker, Jenny C., Christopher Ksoll and Travis J. Lybbert. 2010. "Can Mobile Phones Improve Learning?" *CGD Working Paper.*

Aker, Jenny C., Christopher Ksoll and Travis J. Lybbert. October 2012. "Can Mobile Phones Improve Learning? Evidence from a Field Experiment in Niger." *American Economic Journal: Applied Economics.* Vol 4(4): 94-120.

Aker, Jenny C., Christopher Ksoll, Danielle Miller, Karla Perez, Susan L. Smalley. 2017. "Learning without Teachers? Evidence from a Randomized Experiment of a Mobile Phone-Based Adult Education Program in Los Angeles." *CGD Working Paper 368.*

Aker, Jenny C. and Melita Sawyer. 2018. "Adult Learning in Sub-Saharan Africa: What do and don't we know?" Unpublished working paper.

Andrabi, Tahir, Jishnu Das, Asim Ijaz Khwaja, and Tristan Zajonc. 2011. "Do Value-Added Estimates Add Value? Accounting for Learning Dynamics." *American Economic Journal: Applied Economics*, 3(3): 29–54.

Attanasio, O.P., Kaufmann, K.M., 2014. "Education choices and returns to schooling: mothers' and youths' subjective expectations and their role by gender." *Journal of Development Economics* 109, 203–216.

Banerjee, Abhijit, and Esther Duflo. 2006. "Addressing Absence." Journal of Economic Perspectives 20 (1): 117–32.

Banerjee, Abhijit, Shawn Cole, Esther Duflo and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *The Quarterly Journal of Economics*, 122(3), pp. 1235-64.

Banerji, Rukmini, James Berry and Marc Shotland. 2017. "The Impact of Mother Literacy and Participation Programs on Child Learning: A Randomized Evaluation in India." *American Economic Journal: Applied Economics.* Vol 9, No. 4.

Barrow, Lisa, Lisa Markman and Cecilia Elena Rouse. 2009. "Technology's Edge: The Educational Benefits of Computer-Aided Instruction." *American Economic Journal: Economic Policy*, 1(1), pp. 52-74.

Bengt, Holmstrom. 1979. "Moral Hazard and Observability". *The Bell Journal of Economics*, Vol. 10, No. 1. (Spring, 1979), pp. 74-91.

Bénabou, Roland, and Jean Tirole. 2006. "Incentives and Prosocial Behavior." *American Economic Review*, 96(5): 1652-1678. DOI: 10.1257/aer.96.5.1652

Bengtsson, Niklas and Per Engström. 2013. "Replacing Trust with Control: A Field Test of Motivation Crowd Out Theory**."** *The Economic Journal,* **124.**

Blunch, Niels-Hugo and Claus C. Pörtner. 2011. "Literacy, Skills and Welfare: Effects of Participation in Adult Literacy Programs." *Economic Development and Cultural Change*. Vol. 60, No. 1 (October 2011): 17-66.

Brock, J. Michelle, Andreas Lange, and Ken L. Leonard. 2016. Generosity and Prosocial Behavior in Healthcare Provision: Evidence from the Laboratory and Field. *Journal of Human Resources* 51(1), 133-162.

Bruhn, Miriam, and David McKenzie. 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." *American Economic Journal: Applied Economics*, 1(4): 200-232

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2008. "Bootstrap-based improvements for inference with clustered errors." *Review of Economics and Statistics* 90.3: 414-427.

Carron, G. 1990. "The Functioning and Effects of the Kenya Literacy Program." *African Studies Review*, pp. 97-120.

Cilliers, Jacobus, Ibrahim Kasirye, Clare Leaver, Pieter Serneels, and Andrew Zeitlin. 2018. "Pay for locally monitored performance? A welfare analysis for teacher attendance in Ugandan primary schools."

Cueto, Santiago, Máximo Torero, Juan León, and José Deustua. 2008. Asistencia docente y rendimiento escolar: el caso del programa META "Teacher support and school accountability: The META program". GRADE Working Paper No. 53. Lima, Peru: GRADE.

De Ree, Joppe, Karthik Muralidharan, Menno Pradhan and Halsey Rogers. 2016. "Double for Nothing? Experimental Evidence on the Impact of an Unconditional Salary Increase on Student Performance in Indonesia."

DiNardo, J., J. McCrary, and L. Sanbonmatsu. 2006. "Constructive Proposals for Dealing with Attrition: An Empirical Example." Working paper, University of Michigan**.**

Dubeck, Margaret M. and Amber Gove. 2015. "The early grade reading assessment (EGRA): Its theoretical foundation, purpose, and limitations." *International Journal of Educational Development* 40 (2015) 315–322.

Duflo, Esther, Rema Hanna and Stephen Ryan. 2012. "Incentives Work: Getting Teachers to Come to School," *American Economic Review.*

Duflo, Esther. 2012. "Women Empowerment and Economic Development." *Journal of Economic Literature***.** 50(4). 1051-1079.

Duflo, Esther, Pascaline Dupas and Michael Kremer. 2015."School Governance, Pupil-Teacher- Ratios, and Teacher Incentives: Experimental Evidence from Kenyan Primary Schools". *Journal of Public Economics* Vol. 123, pp. 92-110.

Doepke, Mathias and Michele Tertilt. 2014. "Does Female Empowerment Promote Economic Development?" *NBER Working Paper 19888,* NBER, Inc.

Evans, D. and A. Ghosh. 2008. "Prioritizing Educational Investments in Children in the Developing World." *RAND Labor and Population Working Paper*, **WR 587.**

Frey, Bruno S. and Felix Oberholzer-Gee. 1997. "The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding- Out." *The American Economic Review* Vol. 87, No. 4 (Sep., 1997), pp. 746-755.

Frey, Bruno S., and Reto Jegen. 2001. "Motivation Crowding Theory." *Journal of Economic Surveys*, 15(5): 589–611.

Glewwe, Paul, and Michael Kremer. 2006. "Schools, Teachers, and Education Outcomes in Developing Countries." In *Handbook of the Economics of Education* Volume 2, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, 945–1017. Amsterdam: Elsevier.

Guerrero, Gabriela, Juan Leon, Mayli Zapata & Santiago Cueto. 2013. "Getting teachers back to the classroom. A systematic review on what works to improve teacher attendance in developing countries." *Journal of Development Effectiveness*. 5(4).

Gurol-Urganci, Ipek, Thyra de Jongh, Vlasta Vodopivec-Jamsek, Rifat Atun, Josip Car. 2013. Mobile phone messaging reminders for attendance at healthcare appointments. *Cochrane Database Syst Rev* 2013. Dec 5;(12):CD007458. doi: 10.1002/14651858.CD007458.pub3.

Jensen, Robert. 2010. "The (Perceived) Returns to Education and the Demand for Schooling." *The Quarterly Journal of Economics*, Volume 125, Issue 2, 1 May 2010, Pages 515–548

Kremer, Michael, Edward Miguel and Rebecca Thornton. 2009. "Incentives to Learn." *The Review of Economics and Statistics*. Vol XCI (3).

Lee, David S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects" *The Review of Economic Studies*, 6, 1072-1102.

Mbiti, Isaac. 2016. "The Need for Accountability in Education in Developing Countries." *Journal of Economic Perspectives.* 30(3): 109-132.

Muralidharan, Karthik, Jishnu Das, Alaka Holla and Aakash Mohpal. 2017. "The Fiscal Cost of Weak Governance: Evidence from Teacher Absence in India." *Journal of Public Economics*. 145: 116-135.

Ortega, Daniel and Francisco Rodríguez. 2008. "Freed from Illiteracy? A Closer Look at Venezuela's Mision Robinson Literacy Campaign." *Economic Development and Cultural Change*, 57, pp. 1-30.

Osorio, Felipe, and Leigh L. Linden. 2009. "The use and misuse of computers in education: evidence from a randomized experiment in Colombia." *The World Bank Policy Research Working Paper Series.*

Oxenham, John, Abdoul Hamid Diallo, Anne Ruhweza Katahoire, Anne Petkova-Mwangi and Oumar Sall. 2002. *Skills and Literacy Training for Better Livelihoods: A Review of Approaches and Experiences*. Washington D.C.: World Bank.

Reubens, Andrea. 2009. "Early Grade Mathematics Assessment (EGMA): A Conceptual Framework Based on Mathematics Skills Development in Children." USAID: Washington, D.C.

Romain, R. and L. Armstrong. 1987. *Review of World Bank Operations in Nonformal Education and Training*. World Bank, Education and Training Dept., Policy Division.

Royer, James M., Helen Abadzi and Jules Kinda. 2004. "The impact of Phonological-Awareness and Rapid-Reading Training on The Reading Skills of Adolescent and Adult Neo-Literates." *International Review of Education*. 50(1): 53-71.

Ryan, R. M. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology*, *43*, 450-461.

Sarkar, Siddharth, Priya Sivashankar, and Hiramalini Seshadri. 2015. "Mobile SMS Reminders for Increasing Medication Adherence." *International Journal of Pharmaceutical Sciences Review and Research*, 32(1), May – June 2015; Article No. 38, Pages: 228-237

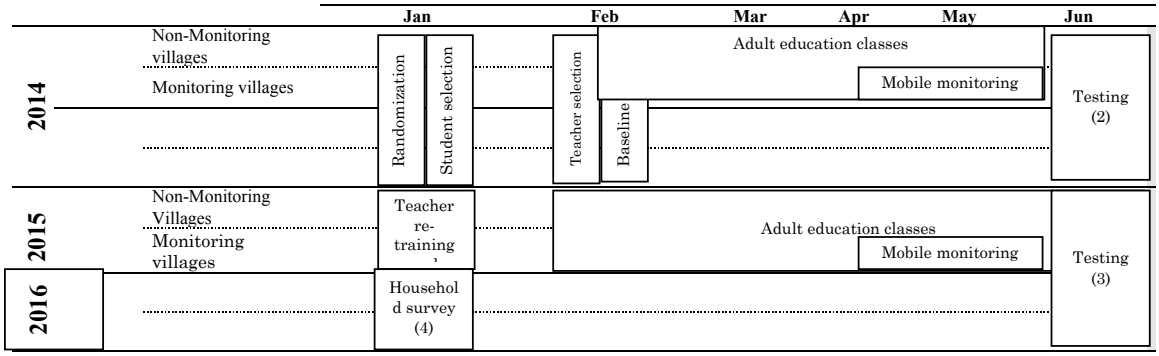
Transparency International. 2013. *The Global Corruption Report: Education.* Routledge: New York, NY.

UNESCO. 2005. *Education for All: Global Monitoring Report. Literacy for Life*. Paris: UNESCO.

UNESCO. 2008. *International Literacy Statistics: A Review of Concepts, Methodology and Current Data.* Montreal: UNESCO Institute for Statistics.

UNESCO. 2012. *Education for All: Global Monitoring Report. Youth and Skills: Putting Education to Work.* Paris: UNESCO.

Figure 1. Timeline of Activities



Note: Figure shows the timeline of activities in our study. Amongst the final 131 villages in our sample, 20 were assigned to pure control villages and 111 were assigned to any adult education courses. Amongst those in the adult education classes, villages were randomly assigned to the standard adult education course or the adult education course plus mobile monitoring treatment.

Figure 2. Map of Villages and Treatment Assignment

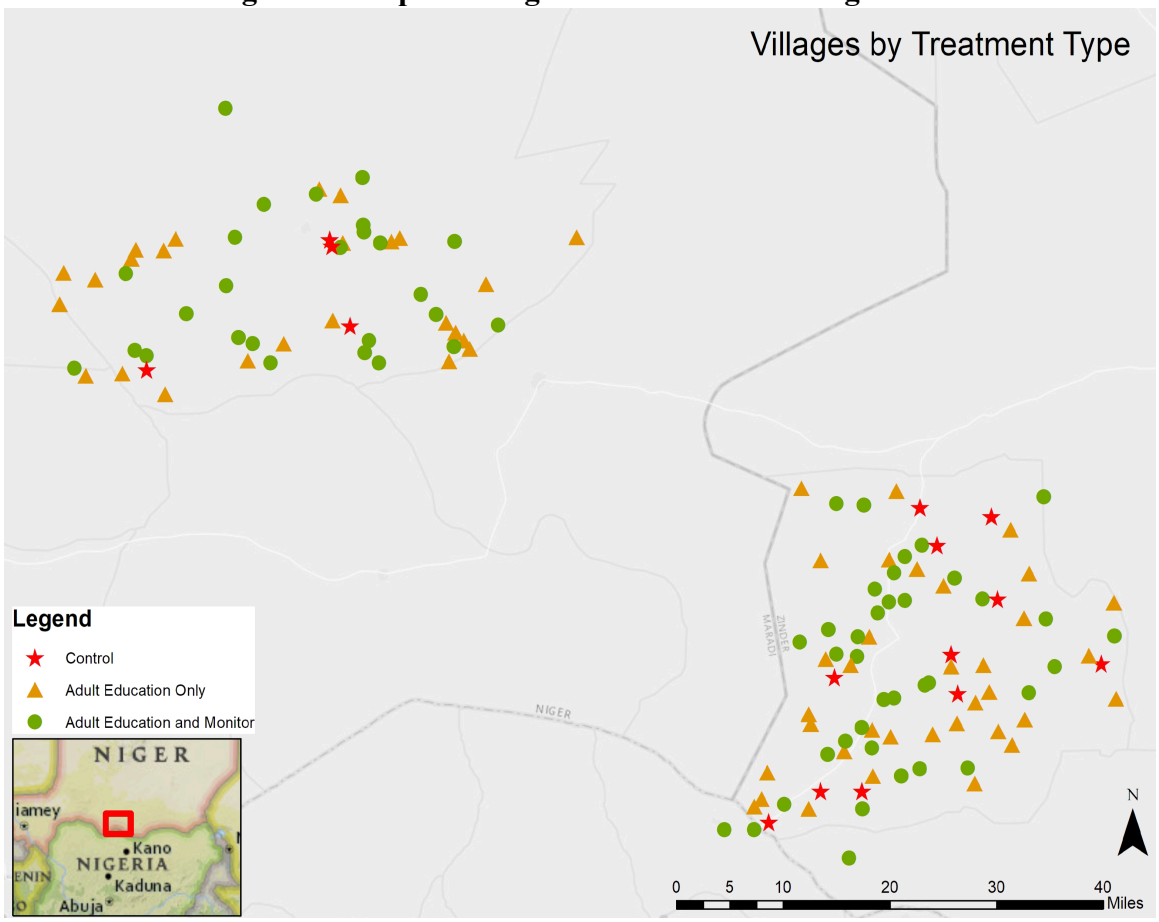


Figure 3. Number of Villages in Each Treatment Group

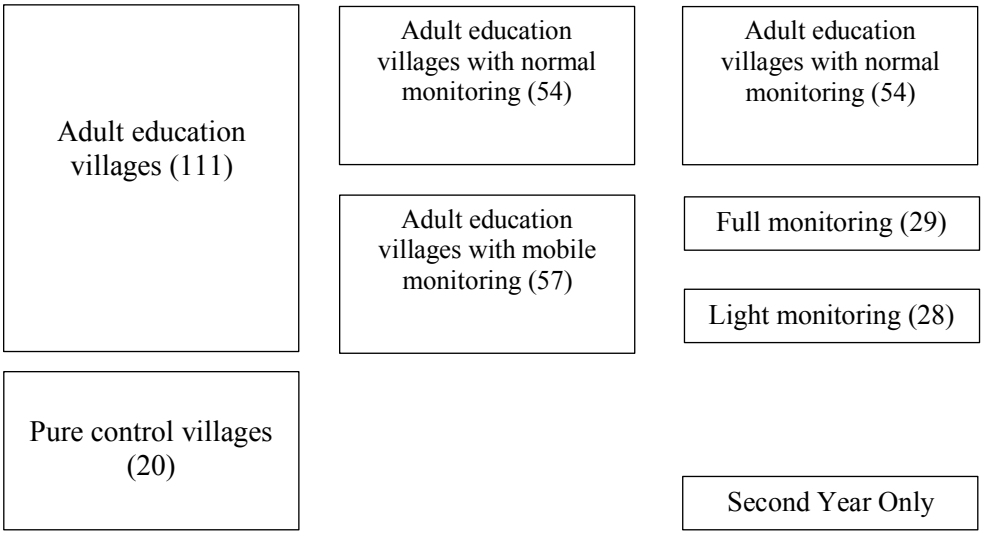
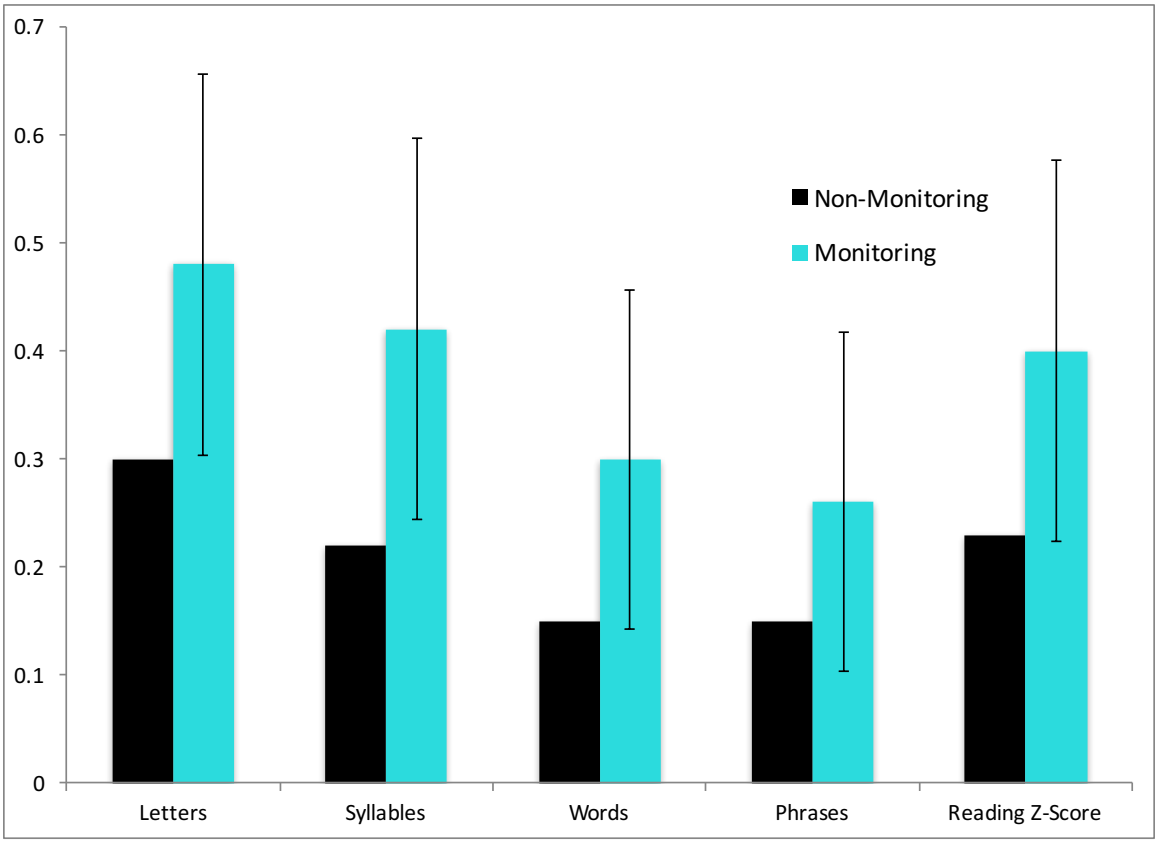


Figure 4: Channels of Potential Impact of the Mobile Monitoring Intervention and Predictions

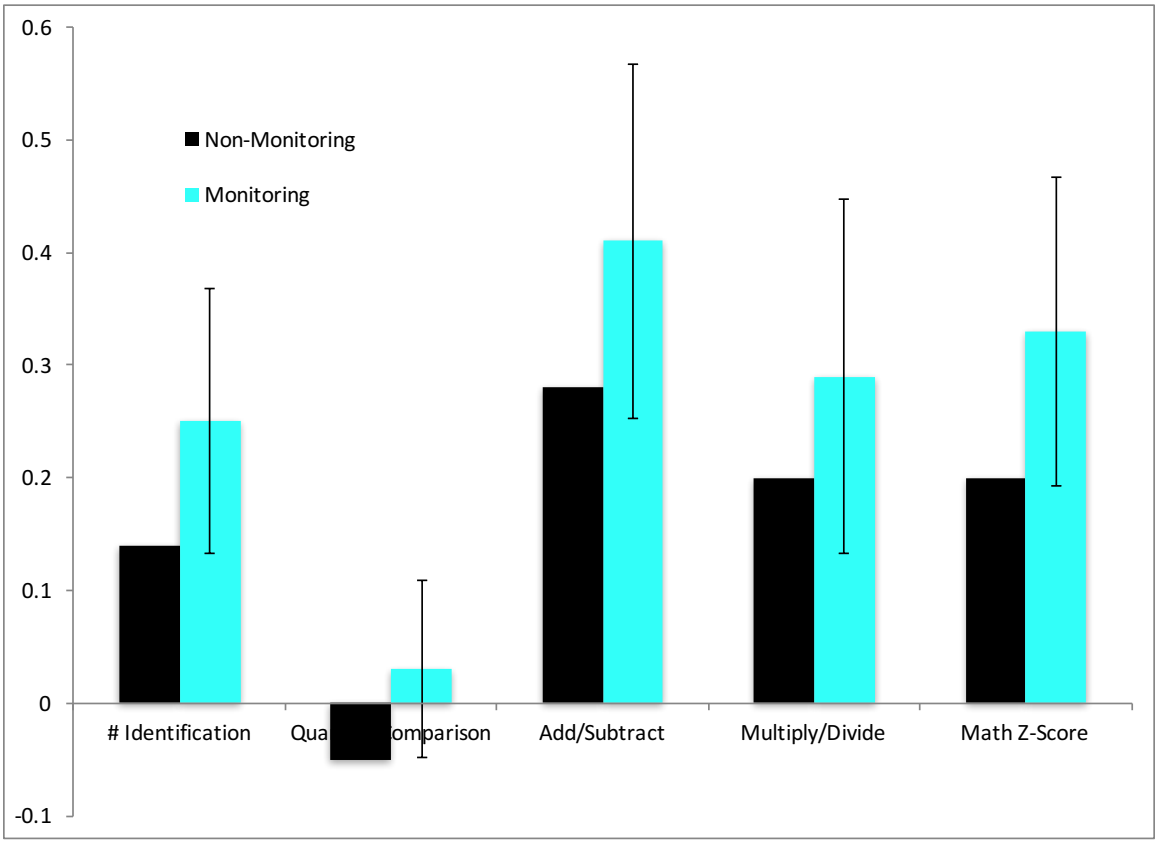
	Prediction for:					
Potential Mechanism	Teacher presence	Teacher preparation	Teacher motivation	Student attendance	Student motivation	Student Outcomes
Teacher effort/shirking	+					+
Teacher's intrinsic motivation	-/+	- /+	-/+			-/+
Teacher's validation of the importance of work	+	+	+			+
Teacher reminder to teach class	+					+
Student effort/shirking				+		
Student perception of the value of adult education				+	+	+
Student reminder				+		+

Figure 5A. Impact of the Monitoring Program over Both Years



Notes: This figure shows the mean timed reading z-scores of different reading tasks for students in monitoring and non-monitoring villages, controlling for stratification fixed effects. Timed reading scores are normalized according to contemporaneous reading scores in control villages. Standard errors are corrected for heteroskedasticity and clustered at the village level.

Figure 5B. Impact of Monitoring on Math Z-Scores over Both Years



Notes: This figure shows the mean math z-scores of different math tasks for students in monitoring and non-monitoring villages, controlling for stratification fixed effects. Math scores are normalized according to contemporaneous math scores in control villages. Standard errors are corrected for heteroskedasticity and clustered at the village level.