Community Monitors vs. Leakage: Experimental Evidence from Afghanistan*

Eli Berman†, Michael Callen‡, Luke N. Condra§, Mitch Downey¶, Tarek Ghani‖, Mohammad Isaqzadeh**

June 14, 2017

Abstract

We study whether trained community-based monitors improve the quality of road construction in Afghanistan, a context of dramatic development spending, pervasive corruption, and a violent conflict preventing government and international workers from monitoring contractors. We partner with a large, well-established organization to randomize its monitor training program, which couples technical training in construction inspections with community mobilization for accountability. Our experiment spans five Afghan provinces and four years of road quality measurement using our own trained technical teams. We find that trained monitors cause dramatic improvements in road quality, producing roads that are better able to endure difficult Afghan winters. Our two-level randomization design shows these effects are not concentrated near trained villages, but spill over to the entire road. Four years after the training, treatment effects have attenuated, potentially because the monitoring program was removed. Finally, interviews with trained monitors suggest the program was effective due to the *combination* of the technical training and the newly created channels of accountability, helping to reconcile mixed findings in past literature where programs often provided only one of these.

Keywords: Community-based monitors, Corruption, Afghanistan, Field experiment

JEL Classification Numbers: C93, H41, O18

*We thank the World Bank and the International Growth Center for funding; Craig McIntosh, Paul Niehaus, Ben Olken, and Jakob Svensson for helpful discussions; and Rachel Berman, Tiffany Chou, Liz Hastings, Wayne Sandholtz, and Erin Troland for valuable research assistance. All mistakes are our own.

†Corresponding author. Department of Economics, UCSD. Email: elib@ucsd.edu

‡Rady School of Management, UCSD. Email: mjcallen@ucsd.edu

§Graduate School of Public and International Affairs, University of Pittsburgh. Email: lcondra@pitt.edu

¶Institute for International Economic Studies, Stockholm University. Email: pmdowney@ucsd.edu

‖Olin Business School, Washington University in St. Louis. Email: tghani@wustl.edu

**Department of Politics, Princeton University. Email: mri2@princeton.edu

"We have to rely on the community to take action; the government is corrupt and they will not sanction contractors."

– Trained community monitor

1 Introduction

Since 2001, the international community has spent $127 billion on relief and reconstruction in Afghanistan (SIGAR, 2017a), a staggering number relative to both the country's GDP ($20 billion in 2015) and other development spending (totalling 10% of development assistance to Sub-Saharan Africa, for instance, despite having only 3% the population). Though this is an extreme case, development spending often targets conflict-prone and politically unstable countries.[1] In the midst of such spending, there is widespread concern of endemic corruption. Transparency International's most recent rankings place Afghanistan as the 8^{th} most corrupt country in the world, an improvement from 2012 when it tied for first. The most recent nationally representative survey of Afghans found that 32% of households had paid a bribe to obtain a state service (estimating that total bribes paid came up to $2 billion in 2014) and found that citizens considered corruption the second most important problem in the country (Isaqzadeh and Kabuli, 2014).[2] Observers have noted that corruption in Afghanistan permeates elections, all levels of government ministries, the courts, banks, and the military (SIGAR, 2016; Transparency International, 2016; USAID, 2009). USAID (2009) summarizes: "The domestic and international consensus is that corruption [in Afghanistan] has become pervasive, entrenched, systemic and by all accounts now unprecedented in scale and reach."

Many worry this scale of corruption will undermine the state's efficacy. Stepping down as Commander of the International Security Assistance Force, General John Allen (2014) warned "the great challenge to Afghanistan's future isn't the Taliban, or the Pakistani safe havens, or even an incipiently hostile Pakistan. The existential threat to the long term viability of modern Afghanistan is corruption," calling the Taliban "an annoyance compared to the scope and magnitude of corruption." These problems interact: In a large series of interviews, Ladbury (2009) found that frustration with corruption was one of the strongest forces driving Afghans to support the Taliban. Not only does corruption threaten to undermine Afghanistan's institutions, but those institutional weaknesses make it more difficult to address corruption. Explaining the challenge of engaging senior leadership in addressing corruption, a senior Afghan National Security Advisor said "corruption is not just a problem for the system of governance in Afghanistan; it is the system of governance" (SIGAR, 2016).

[1] The 2014 top 15 recipients also included Syria, Pakistan, Turkey, Nigeria, and the Democratic Republic of the Congo.

[2] The 2016 National Corruption Survey was not nationally representative due to security concerns.

Contexts where corruption is systemic and all levels of political officials appear involved are those in which bottom-up initiatives seem most promising. One popular form of bottom-up anti-corruption program is community-based monitoring, where citizens themselves seek to hold contractors and government officials directly accountable. However, evaluations of these programs have produced mixed results.[3] This may reflect that "community-based monitoring" can refer to two different mechanisms: information (where citizens learn the quality of service provision) and enforcement (where citizens are given recourse to hold providers accountable). Existing programs vary widely in their relative importance.

We experimentally evaluate a program training community-based monitors which combines both. The program is run by Integrity Watch Afghanistan (IWA), a large organization operating in Afghanistan since 2005. Since its infrastructure monitoring began in 2007, it has trained 1,700 volunteers who have monitored 900 projects, across seven of Afghanistan's 34 provinces (accounting for 26% of its population). The organization enjoys a great deal of respect and influence within Afghan and international civil society organizations. Its biennial National Corruption Survey (run since 2007) helped make corruption a major issue for donors (SIGAR, 2016). Its founder, Dr. Yama Torabi, stepped down as Executive Director in 2014 to join (and later chair) the Independent Joint Anti-Corruption Monitoring and Evaluation Committee, an independent government agency charged with monitoring and evaluating efforts to fight corruption. He was replaced by Sayed Ikram Afzali, who currently simultaneously chairs the Oversight Commission on Access to Information, the agency tasked with improving transparency of and public access to government procurement information.

We evaluate IWA's infrastructure monitoring program in the context of a large road construction initiative. We conduct a large-scale field experiment across five provinces of Afghanistan, accounting for 14% of its population (in Table 1 we show these provinces are broadly representative of other Afghan provinces). We used technical teams to measure road quality over the four years following training. We find that training monitors has no effect on observable road quality measured immediately after construction (one year after training), but that a year later, after the harsh Afghan winter, control roads deteriorated much more dramatically than roads treated with monitors.[4] One year after the construction (two years after the training), roads with treated villages are roughly one standard deviation higher quality. We use a two-level randomization strategy to estimate effects of trained monitors on construction quality in the immediate vicinity of their village, as well as spillover effects elsewhere along the

[3]Specifically, Björkman and Svensson (2009), Duflo, Dupas, and Kremer (2015), and Marcovaldi and Dei Marcovaldi (1999) find positive effects; Banerjee et al. (2010), Lieberman, Posner, and Tsai (2014), and Olken (2007) find no effects.

[4]Roads in our sample are small, and therefore our technical quality inspections were more "surface-level" than those in Olken (2007) and did not, for instance, remove deep core samples.

road. The effects are driven entirely by the share of villages on the road received training, not by whether the village near the measurement received it. In other words, training monitors in one village improves construction quality along the entire road, rather than only near the village. Four years after the initial training, the effects on road quality are similar but smaller and no longer statistically significant. We attribute this to another round of construction in the interim that targeted the lowest quality roads and to IWA scaling back its involvement in some of the treated villages.

The program we evaluate is not a small-scale pilot but a continuously operated, stable design that has been refined over four years of operations by a large, well-funded organization. It is a composite program that includes both types of community-based monitoring interventions. To provide information, it trains monitors on technical features of construction quality (e.g., assessing quality of materials) and improves access to the documents with the original contract specifications. To improve enforcement, it establishes regular meetings between the Afghan government, international donors, and the news media where it updates parties on its monitors' reports and contractors' performance. IWA believes both components are essential, so we did not attempt to violate their program model by implementing only one piece at a time. Instead, we conduct focus groups with IWA-trained monitors and probe them about the importance of these mechanisms. We find that both are important and complementary. In support of the information channel, several monitors reported that they had been doing monitoring before the training but did not know what to look for. In support of the enforcement channel, several monitors reported that they had gone directly to the government in the past, but that it was ineffective until IWA got involved and was able to correct contractor behavior.[5]

Our results contribute to both policy and research. Regarding policy, Afghanistan's 2017 5-year plan sets infrastructure as the top development priority (Islamic Republic of Afghanistan, 2017, p. 14), and a program like IWA's will be important to ensure the effectiveness of those investments. At the same time, IWA has received substantial external funding from international organizations (including the World Bank, the US Institute of Peace, Transparency International, and aid agencies from Germany, the UK, the US, and the EU). IWA's effectiveness in reducing leakage means such funding might be a valuable complement to other development spending. This is particularly important given the view that foreign aid was key in exacerbating Afghan corruption by putting in "too much money, too quickly, into too small an economy, with too little oversight" (SIGAR, 2017b).[6]

[5]This is also consistent with the effect wearing off somewhat as IWA decreased its involvement.

[6]This view is best summarized by former President Karzai saying, "the big corruption, the hundreds of millions of dollars of corruption, it was not Afghan. Now everybody knows that. It was foreign. The contracts, the subcontracts, the blind contracts given to people, money thrown around to buy loyalties, money thrown around to buy submissiveness of Afghan government officials... That was the major part of corruption" (BBC,

This paper relates to the growing literature experimentally evaluating community-based monitors emerging since the 2004 World Development Report focused on "enabling [the poor] to monitor and discipline service providers."[7] We make five contributions to this work.

First, we evaluate community-based monitors in Afghanistan, a conflict-ridden environment. Corruption is endemic in these environments, but states typically lack the capacity to monitor contractors themselves. Importantly, a recent literature shows that successful investment programs can reduce violence and shift popular sentiment from rebels to the government.[8] Thus, the places where reducing leakage is most important are those where top-down solutions are least practical, and it is critical to understand community-based alternatives here.[9]

Second, our experimental design randomizes treatment at the village-level, but also the intensity (share of villages treated) at the road-level.[10] This design, new to the community-based monitoring literature, explicitly allows us to test for spillovers, which we empirically find to be important.

Third, we evaluate a large scale, pre-existing program. Many programs in the literature were created shortly before the evaluation (Lieberman et al. (2014) is an exception). The fact that our program was operating well before our evaluation and that we, as researchers, did not influence its operations or implementation is important for the potential replicability and sustainability of the program and our findings, especially given recent emphasis on challenges in scaling successful pilots (Bold et al., 2016; Muralidharan and Niehaus, 2016).

Fourth, we estimate effects on road quality one, two, and four years after training, a relatively long follow-up in this literature (Björkman and Svensson (2009) call for just this sort of extension). Our results caution reliance on short-run effects. Our primary road quality effects are not present after the initial round of construction (one year post-treatment), but manifest only after a harsh winter has dramatically eroded the control roads (two years post-treatment). Our four-year follow-up (after IWA has scaled back involvement in some districts and another construction initiative targeted the worst roads) then shows somewhat attenuated effects.

Finally, we combine our empirical analysis with qualitative data from interviews of trained

2017). See SIGAR (2016) for a broader review of the role of foreign donors in Afghan corruption.

[7]These evaluations include effects on health (Björkman and Svensson, 2009), education (Banerjee et al., 2010; Duflo et al., 2015; Lieberman et al., 2014), and road construction (Olken, 2007).

[8]See, for instance, Beath, Christia, and Enikolopov (2017) in Afghanistan; Berman, Shapiro, and Felter (2011) in Iraq; and Crost, Felter, and Johnston (2016) in the Philippines. See Berman and Matanock (2015) for a review.

[9]In some ways, however, Afghanistan is an ideal environment for community-based monitors and our results may not translate. In particular, the main barrier to monitoring programs is elite capture, which Björkman and Svensson (2010) and Mansuri and Rao (2013) suggest increases in income inequality. Of the 138 countries with data in the 2013 World Development Indicators, Afghanistan has the 8^{th} lowest Gini coefficient. Björkman and Svensson (2010) and Olken (2006) also find that ethnic heterogeneity undermines monitoring. We collected this data for villages in our sample, and only one road had villages with any appreciable degree of heterogeneity.

[10]See Bobba and Gignoux (2014), Miguel and Kremer (2004), and McIntosh et al. (2014) for similar designs.

monitors to adjudicate between the main two mechanisms identified in the CBM literature: information and enforcement. Our interviews suggest both are important.

In the next section, we describe the IWA intervention. Section 3 describes the design of our evaluation. Section 4 presents results before Section 5 explores possible mechanisms. Section 6 concludes.

2 Intervention

While Afghanistan has seen massive infrastructure investments in recent decades, like many developing countries it struggles with corruption and leakage of public funds is a constant concern threatening to undermine the effectiveness of these investments (SIGAR, 2017a). In this context, we evaluate IWA, which endeavors to improve construction quality and public good provision in places where formal governance is weak. Specifically, IWA works with community-selected volunteers and trains them in a combination of engineering and accounting skills.[11] Their goal is to teach members of the community what technical aspects to look for in evaluating construction quality, as well as how to examine contractors' ledgers to look for leakage. While monitoring is primarily done by the villagers, IWA has engineers on staff to respond to monitors and answer questions or provide additional expertise, as needed.

In addition to training, IWA also establishes semi-formal accountability mechanisms called Provincial Monitoring Boards, which include representatives from the Ministry of Rural Rehabilitation and Development (MRRD), Provincial Councils, IWA-trained monitors, construction contractors, and sometimes aid agencies (including the organizations that funded the road construction) and the media. At these meetings, they discuss construction quality, contractor performance, and potential misappropriation of funds. Finally, IWA emphasizes the importance of *informal* accountability through monitor-led community mobilization. They summarize these dual tasks as "corruption awareness and community mobilization."

An IWA-trained monitor collects project documents, including the technical specifications of the contract. During construction, they visit the site repeatedly, sometimes three times per week. Contractors are aware that these monitors are affiliated with IWA and that IWA engineers are available for the monitors to consult as needed. Monitors first attempt to resolve issues with the contractor directly and have the option of referring them to the Provincial Monitoring Boards when that fails. Below, we present qualitative data from interviews and focus groups to better understand *how* this suite of services improves construction quality.

[11] While we refer to these monitors as volunteers, IWA pays them a small stipend (roughly 15 USD per month), which our focus group (discussed below) describes as too small to make this a profession or make it financially appealing.

3 Empirical strategy

3.1 Experimental design and data collection

Our experiment began in 2011 with a list of villages that IWA was interested in targeting for training monitors. These villages were drawn from five provinces concentrated in the east (see the map in Figure A1).[12] Table 1 shows the characteristics of our sample provinces, relative to those where IWA does not operate (excluding Kabul, an outlier in many dimensions).

Our provinces are broadly similar to others in population and economic structure. They are perhaps slightly more disadvantaged (higher poverty and lower labor force participation, though similar malnutrition, unemployment, and electricity and water access), though better in education and gender equality. The largest differences are in terms of remoteness (our provinces have higher population density and more people living near roads), violence (a third to half the rates of terrorism and Taliban activity), and consequently government presence (higher numbers of government employees and rates of birth certificates). During experiment planning, the government and the World Bank both stated that they could not send their own monitors to these remote villages due to safety concerns. The fact that our provinces are *still* too unsafe for such workers underscores the challenges of monitoring performance in conflict environments and the promise of engaging the local population.

[Table 1 about here.]

Prior to the first wave of road construction (planned for late 2011 and 2012), we conducted baseline surveys of villagers in each of the villages IWA was considering training. The surveys collected villagers' perceptions of road quality, as well as characteristics about the presence of Community Development Councils, women's participation in community decisions, ethnic heterogeneity, etc.

At the same time, we hired technical teams unaffiliated with IWA to measure road quality. We recruited civil engineers with professional training relevant to these measurements and then oversaw their training to assess road safety, the quality of gravel used, and the camber of the road (which determines the risk of erosion). While baseline villager surveys were being conducted, these technical teams were in the field with GPS devices tracking the location of their road measurements. We recruited college-educated engineers, so our teams already possessed the relevant professional skills and our training focused on standardizing practices across teams. Appendix A presents additional details on recruiting and training these monitors.

[12] One road in Herat was also in the experiment, however it is excluded from our main estimation sample for reasons discussed below.

At the time of random assignment of the IWA training program, the villager surveys were available but the technical team assessments were in progress and not yet cleaned. The research team faced substantial time pressure because IWA needed to begin training monitors before the winter made many of these villages inaccessible. Thus, we did assignment based only on villager-assessed road quality, which is correlated with technical team baseline quality at 0.37 (this was unknown at the time of assignment).

For each road, we calculated average quality based on respondents in villages living along the road. We then stratified the sample into terciles to ensure rough balance in baseline quality. Within each tercile, we randomly assigned one-third of *roads* to be pure control, one-third to receive 70% "saturation," and one-third to receive 95% saturation. Next, we randomly assigned villages along these roads to be treated, with the proportion of villages treated equal to the assigned saturation level. That is, if a road was assigned 70% saturation, we randomly assigned 70% of villages on this road to receive monitor training. This design helps us assess spillovers.

Randomization was done according to a "big stick" procedure, where we re-randomized until we could ensure that there were no statistically significant differences (by both village-level treatment and road-level saturation) in several important characteristics (including ethnic heterogeneity, the presence of a Community Development Council, etc.). After randomization, we provided the list of treated villages to IWA for training, in which we took no part.

We conducted three waves of technical team follow-ups (in 2012, 2013, and 2015) to collect road quality measurements and one follow-up village survey (in 2013). The main construction activities occurred between 2011 (our baseline) and 2012, though there was a smaller initiative between 2013 and 2015. Construction contracts and projects were not influenced by treatment: neither the government nor funders were aware of our assignment.

All technical team measurements were conducted by individuals that we trained and paid ourselves, not through IWA. We attempted to send these teams to the same locations that were measured in previous rounds, but this was difficult to implement in practice. Measurements were conducted at frequent intervals along all roads. We aggregate measurements to the village level by taking an average of all technical measurements within a small radius of the village center (our baseline specification uses a 3km buffer and our results are stronger with a 2km buffer, though the sample shrinks).

3.2 Baseline balance

Two features of our experimental design (small numbers of villages per road and big stick randomization) had large effects on the probability that certain villages would be assigned to treatment or control, and that certain roads would be assigned to 0%, 70% or 95% saturation.

These features are described in more detail in Appendix A. Their effects, however, are solely functions of observable baseline characteristics. Thus, we exclude the sample which is not effectively randomized, and random assignment among the remaining sample still allows us to causally estimate the local average treatment effect of trained monitors.

The resulting baseline balance is shown in Table 2, with Panel A displaying balance of village-level assignment and Panel B displaying road-level assigned saturation. The table represents our main estimation sample (that which was effectively randomized), and p-values are based on the clustered (at the road level) wild bootstrap standard errors that we will use throughout. Treated villages and higher saturation roads are slightly worse, though these differences are rarely statistically significant.

[Table 2 about here.]

3.3 Econometric strategy

Our experimental design assigned roads to 0%, 70%, and 95% saturation, and the corresponding share of villages to be treated. However, roads in our main sample only had 2-8 villages (median: 4), so in practice it is rarely possible for the fraction of villages treated to *exactly* match the assigned saturation (because of the integer problem).[13]

As a result, we estimate two types of effects: Intent to Treat (ITT)—based on the reduced form using the assigned saturation—and Treatment on the Treated (TOT)—based on the realized saturation and controlling for the number of villages on the road (which is the reason why assigned and realized saturation do not match). Formally, the ITT estimating equation is:

$$\begin{aligned}\Delta y_{vrt} =& \beta_1 Treat_v + \beta_2 1\{\tilde{S}_r = .7\} + \beta_3 1\{\tilde{S}_r = .95\} \\ &+ \delta_t + \alpha_{q(r)} + \beta_4 y_{vr,t=0} + \beta_5 NumVil_r + \varepsilon_{vrt}\end{aligned} \tag{ITT}$$

where v denotes village, r road, and t time period, y_{vrt} is the road quality measure of road r near village v at time t, and Δy_{vrt} is its difference. The three variables of interest are the village-level treatment assignment ($Treat_v$) and the assigned saturation of the road ($1\{\tilde{S}_r = .7\}$ and $1\{\tilde{S}_r = .95\}$), where $\tilde{S}_r$ denotes the assigned saturation of the road.

In addition to these variables of interest, the specification includes several controls (including time effects δ_t). The $\alpha_{q(r)}$ fixed effects are for the three quantiles of road quality used for stratification in assignment. Because stratification was based only on villager-assessed road quality, we also control for the baseline quality measure under consideration ($y_{vr,t=0}$) to address any (non-significant) baseline imbalance shown in Table 2. Finally, we control for the number

[13] The details of implementation are explained in Appendix A.

of villages on the road ($NumVil_r$), which leads realized saturation to differ from assigned saturation (see Figure A2).

The TOT estimating equation is similar, except that we replace assigned saturation with realized saturation S_r, yielding:

$$\begin{aligned}\Delta y_{vrt} =& \beta_1 Treat_v + \beta_2 S_r \\ &+ \delta_t + \alpha_{q(r)} + \beta_3 y_{vr,t=0} + \beta_4 NumVil_r + \varepsilon_{vrt}\end{aligned} \tag{TOT}$$

In all cases, we use OLS with wild bootstrap standard errors clustered at the road-level to account for correlations at different points along the same road despite our relatively small number of roads (22 in most specifications). When analyzing technical team measurements, we always weight by the number of measurements taken within the buffer.

4 Results

4.1 Graphical evidence

Our main results can be summarized in two sets of figures. First, Figure 1 shows quality over time for both sets of treatment variables. Panel (a) shows that throughout the period, treated (with trained monitors) and control (without) villages had similar road quality, with treated villages having slightly higher quality in 2013.

Panel (b), on the other hand, suggests larger effects. At baseline (2011), roads with more saturation had worse roads. From 2011 to 2012, as construction contracts were executed, both sets of roads experienced similar improvements. A year later in 2013, however, after the harsh Afghan winter, the control roads deteriorated substantially while the more saturated roads remained highly rated in terms of safety. A concern raised in response to an earlier draft was that treatment might have simply delayed construction (through monitoring and auditing) so that treated roads in 2013 look like control roads in 2012. We are skeptical of this concern, as treated roads saw improvements between 2011 and 2012 that were similar to those on control roads. Nonetheless, this partially motivated the collection of 2015 measurements, which show treatment roads did not deteriorate between 2013 and 2015 the way control roads did between 2012 and 2013.

[Figure 1 about here.]

As seen in Table 2 and Figure 1, treatment roads are slightly worse than control roads at baseline, though the difference is not statistically significant. Thus, differences in levels understate differences in growth. To understand the effects of treatment on quality improvements,

and to demonstrate that growth differences are not explained by differences in baseline quality, Figure 2 flexibly shows 2013 quality (Panel (a)) and 2015 quality (Panel (b)) as a function of baseline quality, separately by assigned saturation. Blue squares are villages on roads assigned 70% saturation, purple circles are on roads with 95% saturation, and gray crosses are on control roads. The solid line is a 45-degree line (indicating no change from baseline to follow-up).

[Figure 2 about here.]

It is clear from the figure that at any point in the baseline quality distribution, villages on more saturated roads are higher quality in 2013 than control roads. This result is quite similar in 2015, though somewhat noisier.

4.2 Formal estimates: 2011-2013

Figures 1 and 2 flexibly display the results in a transparent way. The formal estimates of the ITT and TOT specifications are presented in Table 3. The ITT results are in Panel A and the TOT results in Panel B. Both panels use long differences (change from the baseline).

In column 1 we estimate the effects of treatment on the technical teams' improvement in road safety based on 2013 measurements within 3 km of the village. A road being assigned 70% saturation (meaning 70% of villages receive trained monitors) increases road safety by 0.404 on a scale of zero to three. The standard deviation (across all villages and time periods) of this measure is roughly 0.7, so this effect is roughly 60% of a standard deviation and is statistically significant at the 10% level (p-values are in brackets). Being assigned 95% saturation has a larger (0.621) and more significant ($p < .05$) effect. While this shows large effects from the *share* of villages along the entire road that were treated, whether the particular village near the measurement was treated has no appreciable effect (either in magnitude or statistical significance). This implies that trained community monitors have effects throughout the road, not only in the immediate vicinity of their village.

The ITT specification (Panel A) uses the random assignment of saturation. This estimates 70% and 95% saturation independently of one another, and because there are only a finite number of villages on each road, actual and assigned saturation differ. Panel B uses realized saturation (imposing a linear effect), and generates very similar conclusions: road-level saturation has a large and statistically significant ($p < .05$) effect on road safety. A road with 95% saturation is expected to see a 0.794 improvement in road safety,[14] or just over one standard deviation.

[Table 3 about here.]

[14] $0.794 = 0.836 \times 0.95$

This conclusion is robust to a number of alternative specifications. Column 2 restricts analysis to measurements taken within 2 km (rather than 3 km) of the village, which yields effects that are roughly 25% larger and more statistically significant. Column 3 uses an aggregate of three tech team measures taken within the 3 km buffer, where we aggregate by taking a simple sum of standardized z-scores. The results are very similar. Though the ITT estimates are not significant at conventional levels ($p = 0.168$ and $p = 0.104$), the TOT estimates are significant and all coefficients are similar in magnitude to those in column 1.[15] Table B1 shows the results are larger and more statistically significant when failing to control for the baseline quality.[16]

In column 4 we use the villagers' assessment of road quality. We find no effects of treatment on villagers' perceptions, a finding that is consistent throughout. One potential explanation is that villagers are extremely slow to update their perceptions of road quality. Indeed, while villagers' and tech teams' assessments are correlated 0.37 at baseline, the correlation disappears entirely by 2013 (-0.03), consistent with a failure to update accurately.[17]

Finally, Figure 3 shows that the results in column 1 of Panel B are not driven by outliers. We regress both empirical saturation and road safety improvements on the full set of controls and plot the residuals against one another. It is clear that roads with greater saturation systematically experienced greater improvements in road quality (all else equal), an effect which manifests across the full distribution.[18]

[Figure 3 about here.]

4.3 Formal estimates: 2011-2015

Columns 5 and 6 replicate columns 1 and 3 using changes in road safety and the tech team aggregate by 2015. The estimated treatment effects are mostly still positive, though often smaller and never statistically significant. We attribute this to two main forces.

First, there was another wave of construction after 2013, and this construction seems to have targeted the worst roads, which in 2013 were disproportionately control roads.[19] Table B2 in the appendix separately considers quality improvements from 2013 to 2015, and some

[15] Figure B1 in the appendix presents the analog to Figure 2 using this aggregate measure.

[16] Our primary specifications control for baseline quality linearly. Results are unchanged by more flexible controls.

[17] "The challenge with perception-based measures is that they may not measure corruption accurately" (Olken and Pande, 2012, 482). Our finding is consistent with other studies that find a weak correlation between attitudinal and objective measures of corruption (e.g., Olken (2009)).

[18] In addition to these robustness checks, we have estimated effects based on changes from 2011 to 2012. As suggested by Figure 1, we find no significant effects. We have also estimated effects using the full sample (including the villages dropped for reasons discussed in Section 3.2), which yields very similar results to our preferred specification.

[19] Unfortunately, we have no road-level construction data after 2013 to directly assess this.

specifications indeed show that control roads improved by much more than treatment roads did. However, this coefficient is almost entirely eliminated when controlling for 2013 quality. Thus, it seems that largescale investment in control roads (because they were worse) unraveled much of the IWA treatment effect.

Second, though we lack systematic data on IWA engagement, several focus group participants reported that IWA ceased involvement with them after 2013. While some aspects of treatment (e.g., training) may well be long-lasting, others (e.g., gains from being affiliated with IWA and the Provincial Monitoring Boards) require IWA's sustained involvement. If these mechanisms are important (as our focus group suggests is the case) then treatment effects should attenuate when IWA scaled back operations (consistent with Tables 3 and B2).

In summary, the long-run treatment effects of IWA training appear to be positive, but smaller than the medium-run effects and not statistically significant. There are two ways to interpret this result. One is that the intervention is ineffective. Yet an alternative interpretation is that the government can undo the long-run effectiveness of an intervention by spending enough money on replacement construction in the control group. In this case, IWA may not ensure higher quality roads in the long-run, but it does ensure that spending is more effective, one of the major goals of the program.[20]

5 How does IWA work?

How does IWA achieve these gains? The literature on community-based monitoring has emphasized two main channels. The first is *information*: these programs often increase citizens' ability to assess the quality of service provision. The intervention in Björkman and Svensson (2009) is a good example, where it primarily focused on making health service "report cards" publicly available and understood. With road construction this information includes training to evaluate the quality of construction materials being used, the appropriate height of bridge walls, and the angle of the edge of the road (which determines how it will be affected by erosion). IWA trains monitors on these technical aspects, as well as increasing access to the contract specifications to which the road should be built and increasing financial literacy to evaluate

[20] Again, we unfortunately lack post-2013 construction spending data to assess this. We also ruled out three alternative explanations. First, because our original experiment ended in 2013, it is possible that IWA trained monitors in the control villages after this point. We obtained IWA records to assess this, and found that for our main sample, none of the villages on pure control roads were treated after 2013. Second, even without formal IWA training, it is possible that informal monitoring and community mobilization practices spread from one villager to another. In this case, we would expect the control roads nearest to treated villages to experience the greatest 2013-2015 growth. This is not the case. Finally, contractors might be unaware of which villages were treated, and the threat of an IWA-trained village might encourage them to improve quality in any construction project. We obtained a list of contractors awarded projects in the initial round of construction during 2011 and 2012. None of the contractors working in control-intensive districts also worked in treatment-intensive ones.

materials spending and receipts.

The second mechanism is *enforcement*: many programs are coupled with some recourse to pressure service providers. Olken (2007) is a good example, where the monitoring intervention primarily focused on holding village meetings where the contractor needed to account for both the spending and the road quality. Citizens received no training to evaluate performance, and Olken instead tested whether these social pressures and accountability to the community would deter malfeasance. IWA increases enforcement, sanctions, and the ability to hold contractors accountable through its Provincial Monitoring Boards and discussions with government agencies and project funders.

To assess the relative importance of these mechanisms, we conducted a focus group with seven IWA-trained monitors. We posed a series of questions specifically chosen to speak to these mechanisms.[21] Participants felt that both mechanisms were important, and that they complemented one another.[22] Of the various forms of information, monitors most emphasized technical skills and access to the original contract, and of the various forms of enforcement, monitors seemed to find IWA's reputation among contractors most valuable.

5.1 Information

In principle, IWA increases information in three ways: 1) improving citizens' understanding of technical construction quality, 2) enabling scrutiny of contractors' financial records, and 3) increasing access to the contract specifying how the project *should be* executed.

Technical training. All monitors emphasized the value of the training on technical aspects of construction. They acknowledged that the training was fairly rudimentary, and that monitors were "not [as] professional as engineers but at least IWA trained [us] to understand how the quality of the construction materials should be." Some of the monitors even reported that citizens in their village had monitored construction *before* IWA, but that it was ineffective because they didn't know "how to monitor the project quality and implementation."

Contract availability. At the same time, the monitors agreed that the technical training was valuable because they could compare the actual road to the contract. They repeatedly emphasized the importance of having the original contract that the contractor had agreed to, which allowed them to highlight specific shortcomings. Examples included using smaller gravel than was specified and times when the contractor sought to change the path of the road to

[21] In addition to questions about each mechanism, we asked "Which is more important: for communities to have information so that they can engage directly with a contractor, or for them to have advocacy support so that local or central government can engage with a contractor on their behalf?"

[22] Within the existing experimental literature on community-based monitoring, we feel that the most concerted effort to simultaneously combine information and enforcement was Duflo et al. (2015), who also found positive effects.

avoid a particularly rocky section of land. In both cases, monitors made specific reference to the contract. Importantly, this is an IWA treatment effect. Several monitors mentioned that before IWA came to their village, they had requested these contracts in the past, but that contractors refused to provide them. These monitors said that after IWA involvement, contractors became willing to share the documents. Though the monitors agreed that the technical training allowed them to compare roads and contracts, this suggests that increasing contracting transparency might have effects in its own right.

Financial literacy. In contrast, none of the monitors mentioned reviewing the contractors' finances or receipts.

5.2 Enforcement

Information about construction quality is only valuable to the extent that monitors can pressure contractors to change it. IWA seeks to establish several means by which monitors can enforce construction quality: 1) by encouraging and increasing community mobilization, 2) by backing monitors in their engagement with contractors, and 3) by passing complaints to the government.

As with any enforcement, observed instances of a particular method understate its true effects because the threat or possibility of enforcement can deter some bad behavior. Perhaps for this reason, monitors often felt that contractors responded to knowing that someone was watching, and that monitors could get contractors to fix problems simply by asking, without any explicit threats of or references to these modes of recourse. In support that this reflects a deterrent effect, one participant said that monitors derive their influence from the knowledge that the community supports them.

Community. When specific enforcement mechanisms were mentioned, community engagement was the most common. In one village, with the support of the local mullah, the monitors briefed the community about the contractors' performance (both good and bad) during Friday prayers. Other villages used similar "public shaming" approaches.[23]

Government. All monitors were frustrated when describing how the government helps address problems. Contrasting the government and the community, all but one felt the community was more effective.[24] Many pointed to government corruption and collusion with the contractor. Most monitors felt that the government would only take action against a contractor when forced to either by community mobilization or by IWA. In one instance, a large number of

[23]At times, a more extreme form of community mobilization occurred. Villagers would sometimes threaten contractors with violence, and once even bombed a piece of construction equipment.

[24]One case of violence, in which the community destroyed a poorly built building requiring the contractor to redo it, occurred only after district government officials had refused to take action.

villagers gathered outside the district office before they were willing to ask the contractor to redo some construction. In another case, IWA was able to pressure the government into action. The contractor protested that he didn't have enough money for reliable materials because he had to pay so many bribes to government officials. IWA threatened to take the story to the media, and the MRRD got involved and came to an agreement with the contractor.

IWA. IWA's effectiveness in intervening and pressuring contractors was emphasized several times. One monitor said, "IWA name has very high value and all the company understand that is very strong organization." Another one, who had attempted monitoring before IWA, felt that simply having IWA's name connected to him had an impact on contractors' willingness to talk to him.

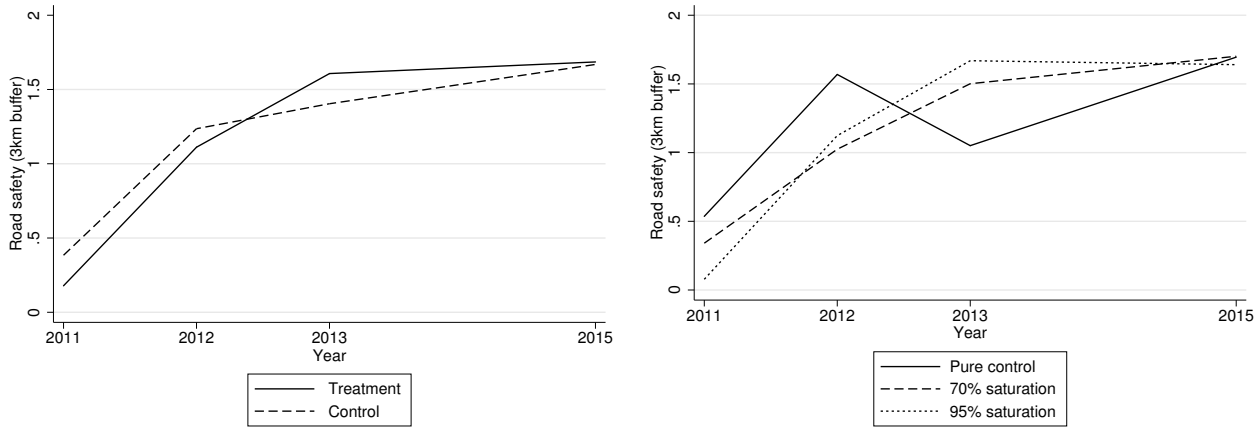
Overall, because of the blend of potential recourse, it seems that IWA-trained monitors are able to improve construction quality often without employing these methods. Ultimately, however, it seems this power is derived from the ability to mobilize the community and the backing of a well-respected organization like IWA.

6 Conclusion

We experimentally evaluated how effective training community monitors can be in improving the performance of construction contractors in rural Afghanistan. This is an environment where state authority is weak and the government and non-governmental organizations have a limited ability to ensure quality. In our context, for instance, the World Bank was unable to send its own auditors to these communities for fear of their safety. Thus, we evaluate whether well-trained citizen volunteers can substitute for the government in situations like these.

We find that training community monitors substantially improves construction quality and contractor performance. These gains are widely shared throughout the road being monitored, rather than only concentrated near the trained villages. The mechanism behind this effect seems to be jointly about information and enforcement. Monitors felt that the skills IWA imparted upon them to assess quality and compare actual construction with the contract were vital, but that it was IWA's reputation and their ability to mobilize their community that allowed them to act upon this information. The results show that in weak states with corrupt or absent government, well-trained citizens can substitute for that government in performing vital monitoring.

Figure 1: Quality over time by assignment status

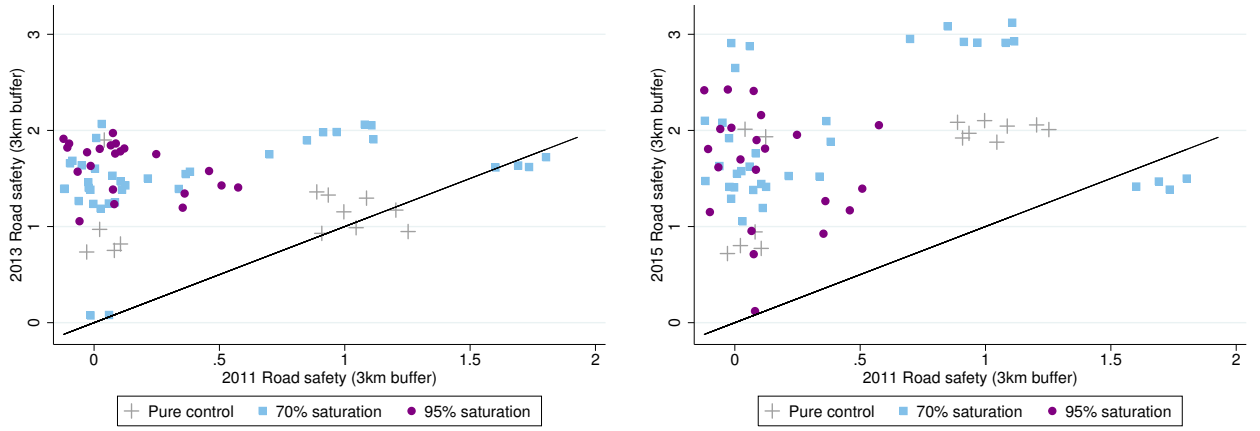


(a) Village-level assignment

(b) Road-level assignment

Both figures based on technical team assessments of road safety in a 3km buffer for sample with baseline technical team measurements available.

Figure 2: Follow-up quality by baseline quality and assignment status



(a) 2013 quality

(b) 2015 quality

Both figures based on technical team assessments of road safety in a 3km buffer. Both figures use a small jitter (a uniform random number on the -.25 to .25 interval) to eliminate overlapping dots.

Figure 3: Non-parametric relationship between treatment and improvements

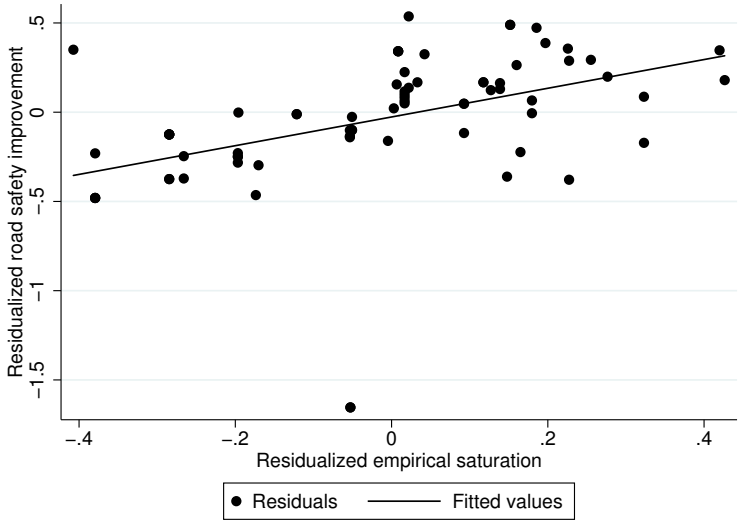


Figure displays relationship between residualized improvements in road safety (2011 to 2013) and residualized empirical saturation (the variable used in the TOT specifications). This is the relationship identifying the saturation coefficient in Table 3, Panel B, column 1.

Table 1: Sample provinces and non-IWA provinces

	Sample IWA Provinces	Non-IWA Provinces (excl. Kabul)
Population and economic activity		
Rural population share	.91	.90
Poverty rate	.45	.35
Caloric deficiency rate	.39	.38
Gini coefficient	25	25.2
Immigrant population share	.073	.048
Labor force participation rate	.50	.54
Unemployment rate	.078	.076
Salary worker share of employment	.20	.11
Agricultural share of employment	.46	.48
Construction share of employment	.048	.083
Education and public services		
Literacy rate	.29	.25
Average years of schooling	2.6	1.9
Girl:Boy primary school attendance ratio	.78	.64
Woman:Man literacy ratio	.46	.13
Share with access to electricity	.61	.65
Share with access to safe drinking water	.40	.36
Share of births with skilled attendant	.40	.34
Remoteness, violence, and government presence		
Population per square kilometer	45	29
Share within 2km of driveable road	.88	.70
Annual terrorism incidents per 100,000 pop.	2.3	4.7
Annual Taliban incidents per 1 million pop.	.9	2.7
Annual terrorism deaths per 1 million pop.	1.9	6.0
Government employees per 100 pop.	3.4	2.5
Share of births with birth certificate	.33	.26
Population	3.4 million	14.5 million

Data drawn from 2011-2012 National Risk and Vulnerability Assessment (Central Statistics Organization, 2014; Ministry of Economy, 2014), except terrorism incidents from the 2005-2009 Worldwide Incidents Tracking System (National Counterterrorism Center, 2010).

Table 2: Baseline balance

	Unconditional means			Conditional difference	
Panel A: Village-level assignment					
	Treatment	Control		Treat. vs. Cont.	
Road safety (3km)	.178	.384		-.156	
Road safety (2km)	.221	.441		-.197	
Tech aggregate (3km)	-1.07	-.838		-.093	
Road quality (vil.)	.847	1.10		.040	
Panel B: Road-level assignment					
	70% Sat.	95% Sat.	Control	70% vs. Cont.	95% vs. Cont.
Road safety (3km)	.341	.078	.536	-.124	-.227
Road safety (2km)	.413	.060	.644	-.419	-.538
Tech aggregate (3km)	-.961	-1.14	-.418	-.372	-.437
Road quality (vil.)	.941	.760	1.49	-.286*	-.264
N. Roads	10	7	5		

* $p < .10$, ** $p < .05$, *** $p < .01$. Distance (3km, 2km) refers to buffer size around village. Road safety and tech aggregate refer to technical team measures (with "road safety" included with two other characteristics in tech aggregate). Road quality refers to village assessment. For all, higher numbers indicate higher quality. Tech aggregates are z-scores (based on the mean and standard deviation across all villages and roads and the full sample period); all other variables are on a 0-3 scale. For road-level assignment, "Sat." indicates assigned saturation. For all, "conditional difference" refers to difference conditional on baseline quality terciles (used in stratification and assignment) and number of villages on road, and stars are based on conditional difference p-values calculated using the wild bootstrap clustered on road (Cameron, Gelbach, and Miller, 2008).Sample size calculations based on the 3km buffer. All estimates based on main sample (see text for details).

Table 3: Community monitors improve construction

	(1)	(2)	(3)	(4)	(5)	(6)
	2011 to 2013				2011 to 2015	
DV: Change in	Road safety, 3k	Road safety, 2k	Tech team agg., 3k	Road quality, vil.	Road safety, 3k	Tech team agg., 3k
	Panel A: Intent to Treat (ITT)					
Village-level treatment	.037	.036	.046	.062	.037	.076
	[.216]	[.408]	[.224]	[.724]	[.744]	[.452]
Assigned 70% saturation	.404*	.529**	.316	-.128	-.067	.003
	[.096]	[.024]	[.168]	[.752]	[.940]	[.952]
Assigned 95% saturation	.621**	.847**	.430	-.245	-.028	.172
	[.024]	[.012]	[.104]	[.404]	[.996]	[.692]
N	72	65	72	86	71	71
	Panel B: Treatment on the Treated (TOT)					
Village-level treatment	.028	.027	.030	.034	-.009	.011
	[.380]	[.548]	[.492]	[.724]	[.956]	[.820]
Empirical saturation	.836**	1.09**	.620*	-.204	.238	.666
	[.040]	[.016]	[.064]	[.552]	[.736]	[.300]
N	72	65	72	86	71	71

* $p < .10$, ** $p < .05$, *** $p < .01$. Displayed in brackets are p-values from wild bootstrap (Cameron, Gelbach, and Miller, 2008) clustered on road quality with 500 replications. "3k" and "2k" refer to 3 and 2 kilometer buffers, respectively, and "agg." refers to an aggregate of three technical team measures (including road safety). All specifications control for terciles of baseline villager-assessed quality (treatment assignment strata), baseline (2011) road quality, and number of villages along the road.

References

Allen, J. R. (2014). Testimony before Senate Subcommittee on Near Eastern and South and Central Asian Affairs, April 30, 2014.

Banerjee, A. V., R. Banerji, E. Duflo, R. Glennerster, and S. Khemani (2010). Pitfalls of participatory programs: Evidence from a randomized evaluation in education in india. *American Economic Journal: Economic Policy*, 1–30.

BBC (2017). Afghanistan's Hamid Karzai says Nato caused 'great suffering'. *Yalda Hakim, 7 October 2013*.

Beath, A., F. Christia, and R. Enikolopov (2017). Direct democracy and resource allocation: Experimental evidence from afghanistan. *Journal of Development Economics 124*, 199–213.

Berman, E. and A. M. Matanock (2015). The empiricists' insurgency. *Annual Review of Political Science 18*, 443–464.

Berman, E., J. N. Shapiro, and J. H. Felter (2011). Can hearts and minds be bought? the economics of counterinsurgency in iraq. *Journal of Political Economy 119*(4), 766–819.

Björkman, M. and J. Svensson (2009). Power to the people: Evidence from a randomized field experiment on community-based monitoring in uganda. *The Quarterly Journal of Economics 124*(2), 735–769.

Björkman, M. and J. Svensson (2010). When is community-based monitoring effective? evidence from a randomized experiment in primary health in uganda. *Journal of the European Economic Association 8*(2-3), 571–581.

Bobba, M. and J. Gignoux (2014). Policy evaluation in the presence of spatial externalities: Reassessing the progresa program.

Bold, T., M. Kimenyi, G. Mwabu, A. Ng'ang'a, and J. Sandefur (2016). Experimental Evidence on Scaling Up Education Reforms in Kenya. *Working Paper*.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics 90*(3), 414–427.

Central Statistics Organization (2014). National Risk and Vulnerability Assessment 2011-2012 Afghanistan Living Condition Survey. *Central Statistics Organization, Islamic Republic of Afghanistan*.

Crost, B., J. H. Felter, and P. B. Johnston (2016). Conditional cash transfers, civil conflict and insurgent influence: Experimental evidence from the philippines. *Journal of Development Economics 118*, 171–182.

Duflo, E., P. Dupas, and M. Kremer (2015). School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from kenyan primary schools. *Journal of Public Economics 123*, 92–110.

Isaqzadeh, M. R. and S. A. Kabuli (2014). Afghan Perceptions and Experience of Corruption. *Integrity Watch Afghanistan National Corruption Survey 2014*.

Islamic Republic of Afghanistan (2017). Afghanistan National Peace and Development Framework: 2017 to 2021.

Ladbury, S. (2009). Testing Hypotheses on Radicalization in Afghanistan: Why Do Men Joint the Taliban and Hizb-i Islami? How Much Do Local Communities Support Them? *Independent Report to the Department of International Development*.

Lieberman, E. S., D. N. Posner, and L. L. Tsai (2014). Does information lead to more active citizenship? evidence from an education intervention in rural kenya. *World Development 60*, 69–83.

Mansuri, G. and V. Rao (2013). *Localizing Development: Does Participation Work?* Washington, DC: The World Bank.

Marcovaldi, M. Â. and G. G. Dei Marcovaldi (1999). Marine turtles of brazil: the history and structure of projeto tamar-ibama. *Biological conservation 91*(1), 35–41.

McIntosh, C., T. Alegría, G. Ordóñez, and R. Zenteno (2014). Infrastructure upgrading and budgeting spillovers: Mexico's habitat experiment. *Working Paper*.

Miguel, E. and M. Kremer (2004). Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica 72*(1), 159–217.

Ministry of Economy (2014). Afghanistan Provincial Briefs. *A joint product of the Islamic Republic of Afghanistan Ministry of Economy and the World Bank*.

Muralidharan, K. and P. Niehaus (2016). Experimentation at Scale. *Working Paper*.

National Counterterrorism Center (2010). Worldwide Incidents Tracking System. *Available from wits.nctc.gov and esoc.princeton.edu*.

Olken, B. A. (2006). Corruption and the costs of redistribution: Micro evidence from indonesia. *Journal of Public Economics 90*, 853–870.

Olken, B. A. (2007). Monitoring corruption: Evidence from a field experiment in indonesia. *Journal of Political Economy 115*(2).

Olken, B. A. (2009). Corruption perceptions vs. corruption reality. *Journal of Public Economics 93*, 950–964.

Olken, B. A. and R. Pande (2012). Corruption in developing countries. *Annual Review of Economics 4*(1), 479–509.

SIGAR (2016). Corruption in Conflict: Lessons from the US Experience in Afghanistan. *Special Inspector General for Afghan Reconstruction Special Report*.

SIGAR (2017a). Special Inspector General for Afghan Reconstruction: April 2017 Quarterly Report to the United States Congress.

SIGAR (2017b). The United States Mission in Afghanistan: A View from SIGAR John Sopko. *Inspector General John F. Sopko speech before Sanford School of Public Policy, Duke University, March 23, 2017*.

Transparency International (2016). From Promises to Action: Navigating Afghanistan's Anti-Corruption Commitments.

USAID (2009). Assessment of Corruption in Afghanistan.

A Implementation Details

[Figure A1 about here.]

A.1 Training technical teams

The technical survey of roads and bridges were conducted by engineers who were trained by the survey firm (ORCA) and overseen by one of the research team members. The candidates had all received formal university training in engineering and had extensive fieldwork experience. The training was delivered by the lead engineer hired by the survey firm while one of the research team members attended the training to make sure the engineers received sufficient guidance and training.

The training was for one day and consisted of three parts: technical review of road and bridge structures and relevant measurements for the technical survey, comprehensive discussion of the technical survey instruments, and logistics and measurement consistency. The first two sections of the training ensured that the engineers had a common understanding of the technical terms in the survey instruments and made measurements of roads and bridges consistently and uniformly. The third component of the training focused on consistency between the follow up round of measurements and the baseline measurements that had been carried out before the intervention. To ensure that the follow up measurements were carried out in the same locations where the baseline measurements had been conducted, the engineers were provided with the list of roads and bridges along with the geospatial markers of the baseline measurements. The geospatial markers of the baseline survey were the only data shared with the engineers and survey firm. The engineers were trained how to use GPS devices to travel to and locate the locations of baseline measurements in the list and to mark those locations in GPS devices after making new measurements. The engineers were evaluated at the end of the training before leaving for the field.

The new geospatial markers recorded by the engineers were later matched with the baseline geospatial markers to confirm that the follow up measurements were done in the same locations as the baseline measurements. In order to verify the consistency and quality of measurements, the data collected by engineers were assessed by the survey firm's quality control team and the lead engineer. Out of the five provinces, the measurements collected in Badakhshan were evaluated to be of low quality and not matching the locations of the baseline measurements. New engineers were trained and new measurements were carried out in that province.

A.2 Implementation issues

In implementing our design, several challenges arose. First, the number of villages along each road was fewer than expected when developing the randomization protocol, in which villages are randomly assigned to treatment up to the point where an additional treated village would exceed the assigned saturation level. As a result, if there is only one village on a road, it is mechanically assigned to control, because assigning it to treatment would result in 100% saturation, which exceeds either the assignment of 70% or 95%. The finite number of villages causes the realized saturation to differ from the assigned saturation in the predictable, deterministic way shown in Figure A2.

[Figure A2 about here.]

Because roads with a single village are deterministically given zero percent realized saturation (irrespective of their randomly assigned saturation), we exclude these roads. To assess the effect of saturation, we use both the assigned saturation (the reduced form) and the realized saturation after controlling for the number of villages (OLS). Results are similar using assigned saturation to instrument for the realization.[25]

Second, our original assignment protocol used a "big stick" procedure to ensure balance along a number of dimensions noted above. Because some of the large roads were outliers in some of these dimensions, big stick randomization essentially forced these roads to receive a certain assignment in order to balance the sample along these dimensions. We re-ran our assignment protocol (with the big stick requirement) 10,000 times and identified three roads that were assigned the same saturation 50% or more of the time, which we exclude (treating them as effectively "not randomly assigned").

Third, unrelated to these issues, our random assignment had relatively poor balance on baseline road quality by random chance (particularly with the road-level assigned saturation variables). There are two types of poor balance. First, despite stratifying on baseline quality tercile in assigning road-level saturation, balance is imperfect. In re-running the randomization protocol, 98% of alternative assignments had better balance between control and 70% saturation and 75% had better balance between control and 95% saturation. Second, the terciles used for stratification in the random assignment were based on villagers' reports of baseline quality and *not* quality as assessed by the technical teams, which is imperfectly correlated (neither measure of quality was used in the "big stick" procedure). In general, our balance seems to be worse in terms of technical team assessments.

[25]Note that a non-linear combination of random assignment and the predetermined number of villages on a road yields a perfect fit for realized saturation.

Finally, many villages in Afghanistan have no formal names or boundaries, and there are often many ways to transliterate village names into the Roman alphabet. Thus, our original assignment of villages to treatment and control included splitting some villages which later data suggested were actually one village. Because IWA was only ever provided lists of treatment villages (not control villages), any village that was ever assigned to treatment was treated. In other words, the "dual assignment" to treatment and control had no effect on whether or how IWA trained the monitors because IWA never knew the village was also assigned to control. Thus, we treat villages assigned to both as being assigned to treatment alone, and the main effect of this is that there are fewer control villages than there would otherwise be.

Figure A1: Sample Provinces

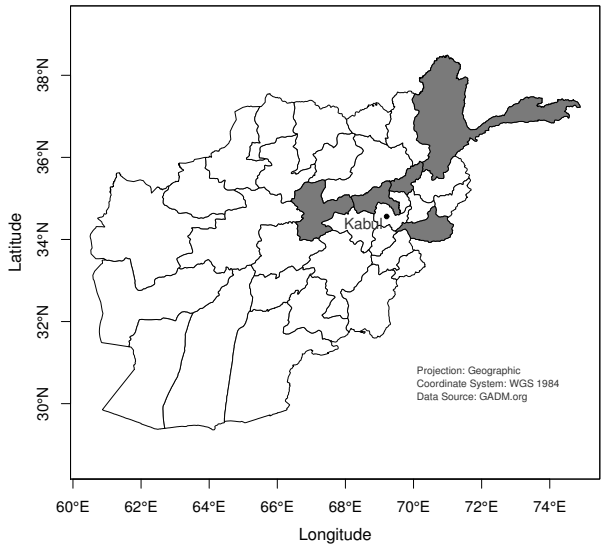


Figure displays sample provinces in gray: Badakhshan, Bamyan, Nangarhar, Panjshir, and Parwan.

Figure A2: Road-level assigned vs. realized saturation

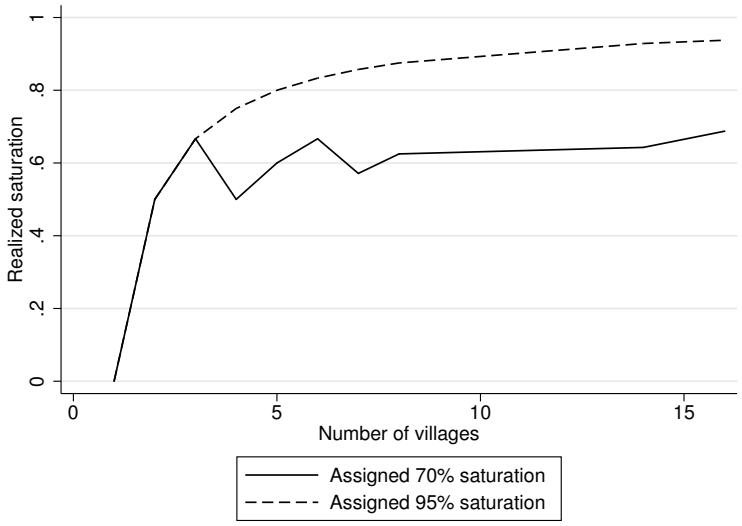


Figure displays relationship between assigned and realized saturation, as a function of the number of villages on the road.

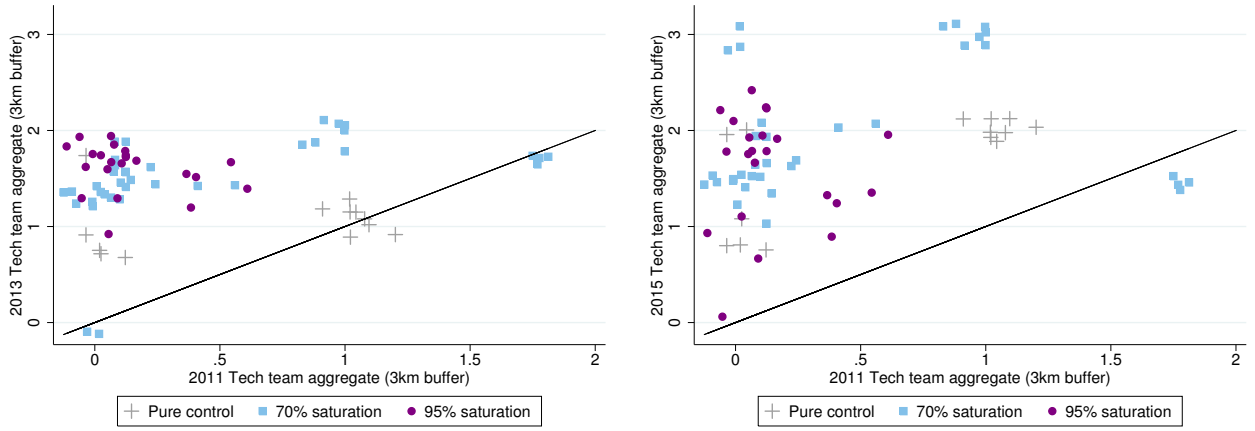
B Additional Results

[Table B1 about here.]

[Figure B1 about here.]

[Table B2 about here.]

Figure B1: Follow-up technical team aggregate by baseline quality and assignment status



(a) 2013 quality

(b) 2015 quality

Both figures based on aggregate technical team quality assessments in a 3km buffer. Both figures use a small jitter (a uniform random number on the -.25 to .25 interval) to eliminate overlapping dots.

Table B1: Main results without controlling for baseline quality

	(1)	(2)	(3)	(4)	(5)
	2011 to 2013			2011 to 2015	
DV: Change in	Road safety	Tech team aggregate	Road quality	Road safety	Tech team aggregate
	Panel A: Intent to Treat (ITT)				
Village-level treatment	.146**	.130*	.130	.163*	.146*
	[.040]	[.092]	[.532]	[.096]	[.100]
Assigned 70% saturation	.640	.842**	.055	.183	.527
	[.112]	[.036]	[.856]	[.676]	[.152]
Assigned 95% saturation	1.03**	1.10**	-.197	.364	.837*
	[.028]	[.012]	[.720]	[.360]	[.064]
N	72	72	86	71	71
	Panel B: Treatment on the Treated (TOT)				
Village-level treatment	.152**	.145**	.011	.131	.110
	[.024]	[.020]	[.920]	[.220]	[.264]
Empirical saturation	1.33**	1.31***	.197	.666	1.31**
	[.040]	[.004]	[.696]	[.200]	[.028]
N	72	72	86	71	71

* $p < .10$, ** $p < .05$, *** $p < .01$. Displayed in brackets are p-values from wild bootstrap (Cameron, Gelbach, and Miller, 2008) clustered on road quality with 500 replications. Road safety and the tech team aggregate are based on measurements in a 3km buffer; road quality is based on villagers' assessments. All specifications control for terciles of baseline villager-assessed quality (treatment assignment strata) and number of villages along the road.

Table B2: 2013-2015 growth differences are explained by 2011-2013 treatment effects

	(1)	(2)	(3)	(4)	(5)	(6)
	2011 to 2013		2013 to 2015			
DV: Change in	Road safety	Tech team aggregate	Road safety	Tech team aggregate	Road safety	Tech team aggregate
Village-level treatment	.028	.030	-.148	-.095	-.051	-.062
	[.392]	[.092]	[.684]	[.792]	[.732]	[.484]
Empirical saturation	.836**	.620*	-.637	-.031	.038	.178
	[.024]	[.092]	[.284]	[.948]	[1.00]	[.680]
2013 road quality					-.758	-.295
					[.732]	[.804]
N	72	72	69	69	69	69

* $p < .10$, ** $p < .05$, *** $p < .01$. Displayed in brackets are p-values from wild bootstrap (Cameron, Gelbach, and Miller, 2008) clustered on road quality with 500 replications. All measurements based on a 3km buffer around the village. All specifications control for terciles of baseline villager-assessed quality (treatment assignment strata), baseline (2011) road quality, and number of villages along the road. Specification is TOT.