Journal Pre-proof

Transparency, governance, and water and sanitation: Experimental evidence from schools in rural Bangladesh

Umrbek Allakulov, Serena Cocciolo, Binayak Das, Md. Ahasan Habib, Lovisa Rambjer, Anna Tompsett

PII: S0304-3878(23)00037-8
DOI: https://doi.org/10.1016/j.jdeveco.2023.103082
Reference: DEVEC 103082

To appear in: *Journal of Development Economics*

Received date : 30 November 2021
Revised date : 23 January 2023
Accepted date : 2 March 2023

Please cite this article as: U. Allakulov, S. Cocciolo, B. Das et al., Transparency, governance, and water and sanitation: Experimental evidence from schools in rural Bangladesh. *Journal of Development Economics* (2023), doi: https://doi.org/10.1016/j.jdeveco.2023.103082.

Transparency, governance, and water and sanitation: experimental evidence from schools in rural Bangladesh

Umrbek Allakulov[1], Serena Cocciolo[2], Binayak Das[1], Md. Ahasan Habib[3,4], Lovisa Rambjer[5], and Anna Tompsett[*6,7]

[1]Water Integrity Network, Berlin, Germany
[2]World Bank, Washington D.C., USA
[3]KTH Royal Institute of Technology, Stockholm, Sweden
[4]NGO Forum for Public Health, Dhaka, Bangladesh
[5]Department of Economics, Uppsala University, Sweden
[6]Institute for International Economic Studies, Stockholm University, Sweden
[7]Beijer Institute of Ecological Economics, Royal Swedish Academy of Sciences, Sweden

[*]**Correspondence: anna.tompsett@iies.su.se.** We thank Merve Demirel, Abdul Haque, Urja Jain, Md. Mustafizur Rahman, and Mir Abu Raihan for exemplary research assistance and support. We also thank the local implementing NGO, Development Organization of the Rural Poor (DORP), for their collaboration and support in the field. Thanks also to Jared Gars, Molly Lipscomb, Andreas Menzel, and seminar participants at Uppsala University and the ADBI-BMGF-IFS conference on Sanitation and Development for helpful comments and suggestions. Pre-registration: AEARCTR-0003111 (https://www.socialscienceregistry.org/trials/3111). Conflicts of interest: The intervention we evaluate was designed by the Water Integrity Network (WIN), and data collection for this study was funded by WIN. Allakulov and Das declare that they are employed by WIN. Habib is employed by NGO Forum for Public Health, which has received funding from WIN, including for the data collection for this study. Cocciolo, Rambjer, and Tompsett have received no funding from WIN and declare no conflict of interest.

Abstract

Can transparency interventions improve WASH service provision? We use a randomized experiment to evaluate the impacts of a transparency intervention, a deliberative multi-stakeholder workshop initiated with a community scorecard exercise, in schools in rural Bangladesh. To measure impacts, we combine survey data, direct observations, and administrative data. The intervention leads to moderate but consistent improvements in knowledge of WASH standards and practices, and institutions for WASH service management, but does not improve school WASH service provision or change WASH facility use patterns. Drawing on rich descriptive data, we suggest several reasons why the intervention we evaluate did not improve WASH service outcomes and propose ways to improve the design of future interventions.

1 Introduction

A common belief among policy-makers is that improving water, sanitation, and hygiene—or WASH—services is largely a question of improving governance. For example, the OECD (2015) argues that crises in water management are "primarily" crises in governance, and the UNDP (2011) identifies corruption—a governance failure—as a major impediment to improving WASH service provision. Many policy-makers thus anticipate that interventions to improve governance should improve service provision in the WASH sector. One popular class of such interventions aims to increase transparency in service provision. In practice, however, previous experimental evaluations of similar interventions in other sectors have yielded only mixed results (Kosack and Fung, 2014).

Using a randomized experiment focused on WASH services in 60 schools in rural Bangladesh, we evaluate the impacts of a transparency intervention comprising a deliberative multi-stakeholder workshop initiated with a community scorecard exercise. Following a pre-specified analysis plan, we trace out the full causal path via which the intervention could affect outcomes of interest: changes in knowledge and institutions, improvements to WASH facilities, changes in use patterns in WASH facilities, school attendance, and exam results.[1] We find that the intervention leads to moderate improvements in knowledge and institutions but not to measurable improvements in other outcomes.

Many schoolchildren in Bangladesh lack access to adequate WASH facilities, potentially increasing water-borne disease and absenteeism (UNICEF, 2012). Before our intervention, two schools in this study had no functioning toilets available to students. Those schools that did have functioning toilets had only one for every 226 students, far below the internationally recommended minimum standard of 1:50 (Water Aid, 2016). Only 61% of functioning toilets were clean and could be locked. No school had any toilet meeting the minimum requirements for safe and private menstrual hygiene management (MHM). Students and teachers reported that inadequate sanitation facilities affected female student attendance.

We randomly assigned 30 schools to receive the intervention we study. Developed by the Water Integrity Network (WIN) and the IRC International Water and Sanitation Centre, the Annotated Water Integrity Scan (AWIS) is a participatory assessment tool designed to gather and disseminate information, improve governance, and promote integrity.[2] Integrity is defined as a set of practices that impede corruption and promote respect for the rule of law. During a workshop, participants first anonymously score WASH service provision across indicators measuring transparency, accountability, and participation. Facilitators then reveal the anonymized scores. When scores differ across participants, facilitators invite them to discuss why, with the goal of understanding differences in opinions rather than defending

[1] The pre-analysis plan that we follow is publicly available from the AEA RCT registry: `https://www.socialscienceregistry.org/versions/54056/docs/version`.

[2] See `https://www.waterintegritynetwork.net` and `https://www.ircwash.org/home`.

individual positions. The output of this stage of the discussion is a single, unified set of scores. This process is designed to yield a common understanding of challenges with the goal of identifying concrete and specific priority actions for improvement during the last phase of the workshop. AWIS closely resembles one of a set of "exemplary transparency interventions" identified by Kosack and Fung (2014).[3]

We evaluate the impacts of the AWIS intervention on three families of outcomes, eight to twelve months after implementation. First, we evaluate effects on knowledge about WASH and MHM practices among male and female students, male and female teachers, and head-teachers, and institutions for WASH service management. For five out of six outcomes, we see small to medium improvements (between 0.1 and 0.3 standard deviations). Not all the individual effects are statistically significant and none independently survive corrections for multiple hypothesis testing that control the family-wise error rate. However, this pattern of consistently positive results is unlikely to have occurred due to chance ($p = .031$).

Second, we turn to material effects on WASH provision and use. We measure effects on three outcomes: the number of functioning toilets available to students; a toilet quality index, which summarizes measures of functionality, cleanliness, and hygiene; and latrine use. We develop a novel approach to anonymously measuring latrine use, using discreet magnetic door sensors that record when the latrine door is opened or closed. The characteristic signature of a latrine use event in these data is an interval during which the door is closed bookended by two intervals during which the door is open. Door sensors present a potential improvement over other alternatives such as motion sensor cameras, because the entire equipment can be installed outside the cubicle, minimizing intrusion of privacy. We estimate effects on WASH provision and use separately for toilets used by male and female students. We also report the number of functioning toilets available to all students (for which use is not gender-separated), although few schools have them. Three out of seven metrics show statistically insignificant improvements, and four show declines. We thus find no evidence that the intervention improved WASH service provision or increased WASH facility use.

Most disappointing, the impacts on WASH provision and use are negative for all metrics for toilets available to female students. We did not design our experiment to detect negative effects of the intervention; we prespecified one-sided hypothesis tests to maximize power, given the limited sample size. However, two out of three of these negative effects are statistically significant at the 10 percent level in two-sided hypothesis tests, a pattern of results with a low likelihood of occurring if the intervention had no effect on school WASH services.

[3]Specifically, Kosack and Fung (2014) identify "citizen report cards or community score cards followed by deliberative community meetings" as the exemplary transparency intervention in contexts in which service providers are willing to cooperate in reforms and the main role of transparency interventions is to feed collaborative problem solving. While it is often uncertain whether or not service providers are willing to cooperate in reforms, and transparency interventions are less likely to be successful when they are not, successful transparency interventions often reveal service providers to be willing to engage in reform (Kosack and Fung, 2014).

WASH provision and use for female students decline in treated schools relative to control schools not because WASH service provision worsens in treated schools but because control schools more successfully improve WASH service provision. Future research should investigate whether similar interventions can have negative unforeseen discouragement effects, for example by creating expectations of future external material support that disincentivize independent local action.

Finally, we turn to effects on school attendance and exam results. We do not find that the intervention leads to measurable changes in attendance or achievement in exams. The 90% confidence intervals exclude increases in female student attendance larger than 4.1 percentage points (p.p) or declines larger than 2.9 p.p. and increases in female student exam results of greater than 0.16 points on an 5 point scale or declines larger than 0.12 points. For female students, baseline mean attendance is 66% and the mean exam score is 2.3.[4] That the intervention does not measurably affect attendance or results may not be surprising, given the mixed results on knowledge and institutions, and WASH provision and use.

Although the number of schools in the study is relatively small, we maximize our ability to detect effects by collecting rich data at a more granular unit or temporal level wherever feasible. Conversely, we minimize the risk of drawing erroneous conclusions by using randomization-based inference and reporting conservative multiple hypothesis tests that control the risk of erroneously rejecting any of the null hypotheses within a family of outcomes. Ex post minimum detectable effects using measures of the variance of the estimated coefficients (that are independent of the realized point estimates themselves) suggest that the study was powered (at significance level 10% and power 80%) to detect increases in indexes of between 0.2 and 0.4 standard deviations, typically characterized as small to medium-sized effects (Cohen, 1988). Thus the reason we do not detect effects on WASH service provision is not solely a consequence of power. Adherence to the study protocol was also high, with all schools assigned to treatment successfully treated.

Our results contribute to the literature on interventions designed to improve public services through greater transparency or accountability. The intervention we evaluate belongs to a class of interventions called "transparency for accountability" that help users of a public service access information that increases their ability to influence service provision for the better (Kosack and Fung, 2014). These reforms are increasingly "in vogue" (Kosack and Fung, 2014), along with a closely-related group of interventions called "social accountability" interventions that attempt to improve service provision through citizen engagement (Fox, 2015). In practice, transparency for accountability and social accountability interventions commonly share a twin focus on "information"—in our case, the scorecard exercise—and

[4]For male students, the 90% confidence intervals rule out increases in attendance larger than 5.3 percentage points or declines larger than 3.3 p.p., and increases in student exam results of greater than 0.07 points on an 5 point scale or declines larger than 0.19 points. Baseline mean attendance is 57% and mean exam score is 2.3.

"participation"—in our case, the deliberative multi-stakeholder workshop. Total spending on transparency and accountability is difficult to measure, but it most likely exceeds US$200 million a year.[5] The enthusiasm from policy-makers contrasts, however, with the mixed experimental evidence regarding the efficacy of these interventions (Fox, 2015; Joshi, 2010; Kosack and Fung, 2014).

Our primary contribution to this literature is to provide the first experimental evaluation of a transparency intervention in the water and sanitation sector, a sector in which many policy-makers expected high impact. Our results show that a single, time-limited transparency intervention improved knowledge and institutions but not WASH service provision. While improvements to knowledge and institutions are intrinsically important, they would also, all else equal, would be expected to improve service provision. We draw on rich qualitative and quantitative descriptive data to understand why improvements in knowledge and institutions did not translate into improvements in WASH service provision in this case. This exercise, in turn, sheds light on results from previous literature and suggests ways to improve the design of future interventions. We identify three potential reasons.

First, we study a relatively "light-touch" intervention: a one-off workshop that draws only on information available to the community. Such interventions are naturally of particular policy interest because they are relatively cheap to implement and would thus be highly cost-effective if they led to meaningful changes. However, time-limited interventions may be insufficient to change outcomes. In particular, the AWIS intervention appears to have more successfully created consensus over problem diagnosis than plans for action. A longer time horizon might allow both to receive equal attention. In previous studies, less effective interventions were often characterized by shorter time horizons (e.g., Banerjee et al., 2010; BenYishay et al., 2022) while more effective interventions were characterized by longer time horizons, more extended engagement, or accountability measures to incentivize adherence to agreed-upon action plans (e.g., Barr et al., 2012; Björkman and Svensson, 2009; Mohanan et al., 2020). Such modifications most likely increase implementation costs, but they may be necessary if transparency interventions are to be effectively deployed.

Second, it is possible that the intervention may have inadvertently created misleading expectations of forthcoming financial support, discouraging stakeholders from seeking independent solutions to their problems. While our study is to our knowledge unique in documenting negative effects of a transparency intervention—though we cannot rule out the possibility that negative effects have been encountered in unpublished work—null effects are common and other studies also note that expectations about development projects "work" could influence how participants behave during similar interventions (Arkedis et al., 2021).

Third, the results may reflect the challenges of scale in intervention design. Schools

[5]Members of the Transparency and Accountability Initiative, a consortium of eight private and public funders, report spending US$200 million annually on transparency and accountability work in the Global South (Fox, 2015).

in rural Bangladesh may not be able to solve all their WASH service problems internally. Funding new infrastructure typically requires schools to secure additional funding from local communities, non-governmental organizations, or, on an ad hoc basis, from government bodies. Local government administrators have limited and fragmented responsibility for secondary school WASH service provision and proved unwilling or unable to attend school-level workshops. Transparency interventions may be more effective if they diversify their approaches to engagement with different stakeholders or simultaneously provide financial support that expands the choices available to stakeholders and frees them from dependence on external funding (as, for example, in successful precedents such as Pradhan et al., 2014).

Our results leave open the possibility that the improvements we observe in knowledge and institutions might translate into improvements in WASH service provision over longer time horizons. Few previous studies evaluate how the effects of transparency or accountability interventions vary over time and those that do tend to be following up successful interventions (Christensen et al., 2021; Nyqvist et al., 2017), implying that their results may not necessarily generalize. Our findings also do not speak to whether or not similar interventions might be more effective in different WASH service contexts. In particular, similar interventions might be more effective when WASH service provision is the core objective of the provider, as with a water utility, rather than a secondary or indirect objective, as it is in schools. Future research may wish to investigate impacts over longer time horizons or in different WASH service provision contexts.

The paper proceeds as follows: Section 2 describes WASH service provision in the study schools at baseline. Section 3 provides more details about the AWIS intervention. Section 4 outlines a theory of change for how the intervention might alter outcomes. Section 5 describes the data, and section 6, the experiment. Section 7 outlines our approach to analysis, and section 8 reports the results. Section 9 summarizes descriptive evidence that suggests reasons why the intervention did not improve WASH service provision. Section 10 concludes.

2 School WASH service provision in rural Bangladesh

School WASH provision in Bangladesh remains very far from adequate. Although between 2013 and 2018 the provision of toilets for secondary school students increased from one toilet per 200 students (ICRRDB et al., 2014) to one toilet per 115 students (Bangladesh Bureau of Statistics, 2020), the latter figure is still less than half the internationally recommended level. Fewer than half of schools had improved, accessible, gender-separated toilets with soap and water available (Bangladesh Bureau of Statistics, 2020). Rural areas lagged behind urban areas (Bangladesh Bureau of Statistics, 2020).

Secondary school governance and management in Bangladesh is largely decentralized and institutionally fragmented (Water Aid, 2016). Secondary schools are regulated by the Directorate of Secondary and Higher Education (DSHE) but few (around 2% nationwide)

are directly managed by the government (Water Aid, 2016). In non-governmental schools, the primary governance structure is the school management committee, a committee formed by local dignitaries, teachers, and parents or guardians. An upazila officer of the DSHE is nominally part of each school management committee but rarely participates actively in school governance, since the upazila-level DSHE offices have only a few staff members that are responsible for overseeing all secondary education in the upazila. The government subsidizes teacher and headteacher wages. Additionally, the government pays tuition for female students as long as they meet attendance and achievement goals, and remain unmarried.

The school management committee is responsible for providing adequate WASH services (Water Aid, 2016) but receives no direct or systematic governmental support to meet this responsibility. The Ministry of Education issued a national circular in 2015 that directs secondary schools to improve WASH services, which in principle is enforceable, but no resources are made available to schools to comply with the circular (Water Aid, 2016). The Department of Public Health and Engineering (DPHE) has a general responsibility for sanitation in rural areas but no specific budget or responsibility for secondary schools. The most decentralized level of local government, the union parishad or council, plays no direct role in governing secondary education. To fund any project, such as investments in construction, repair, or maintenance of school infrastructure, non-governmental schools must primarily raise funds themselves on a case-by-case basis.

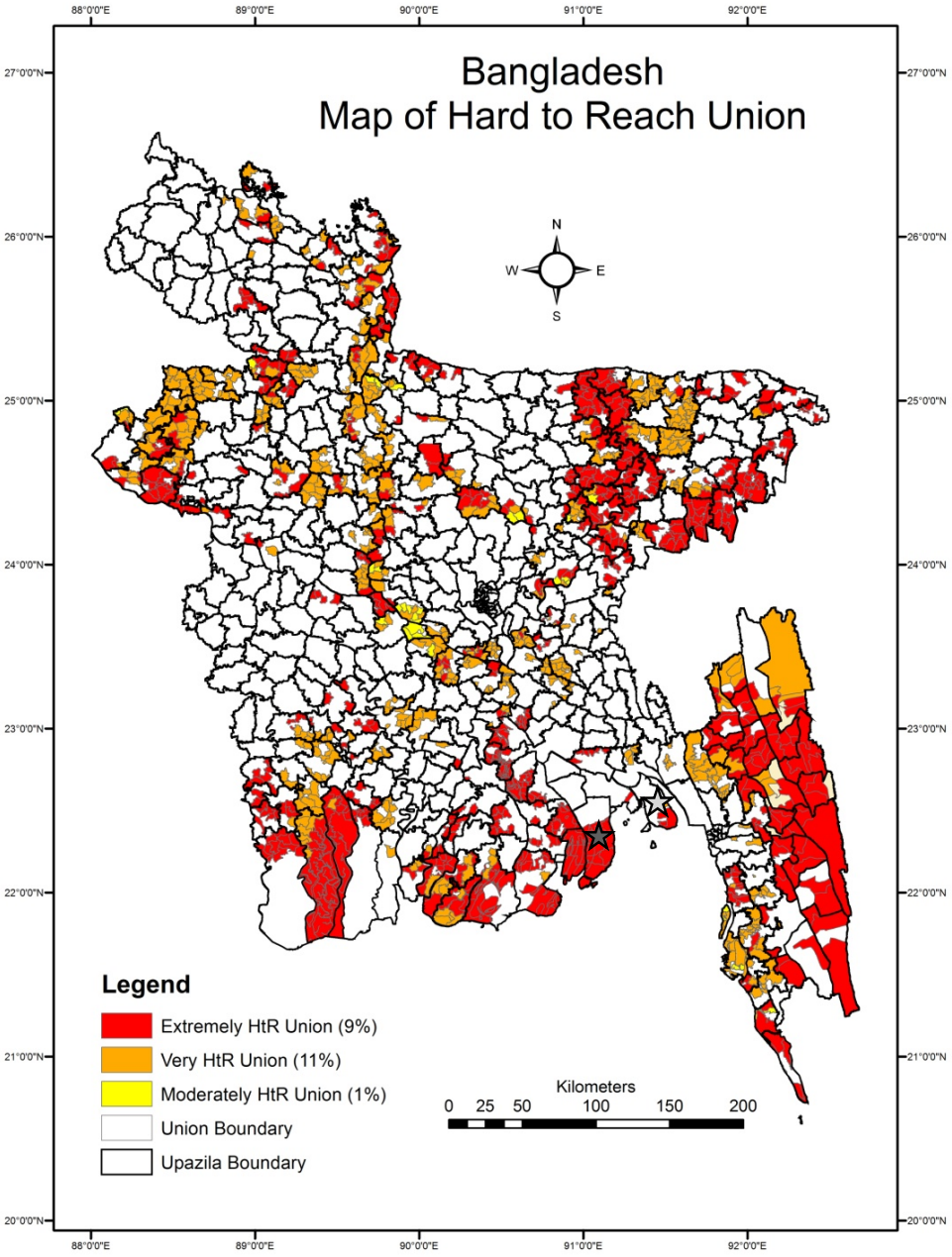
We focus on secondary schools, covering grades 6 to 10, in two *upazilas* (subdistricts), Bhola Sadar and Ramgati, in southern Bangladesh. Both upazilas have a large number of unions (local government units) that are classified as "Extremely Hard to Reach", a classification the Bangladeshi government uses to describe areas with poor water and sanitation coverage and particular obstacles to improvement, such as unfavourable hydrogeological conditions, inadequate communications networks, or frequent natural disasters (see Figure 1).

A local implementing NGO, Development Organization of the Rural Poor (DORP),[6] identified 60 schools with sanitation facilities below national standards to target for the intervention. In practice, almost all schools in the study areas were eligible under these criteria. Since school sanitation facilities may be most important for female adolescent students, we focused on schools with either all female students (11 schools) or both female and male students (49 schools). No school declined to participate in the study, and we did not exclude any schools from the selected pool. All study schools were initially non-governmental schools.[7] Around a third are madrasahs, or Islamic schools. The madrasahs in our study area closely resemble other non-governmental schools in organization and management but offer religious education and Arabic alongside the standard curriculum. At baseline, on average, female-only schools had 424 students and mixed schools had 609 students, of which 318 were

[6]See `https://www.dorpbd.org`.

[7]One study school was nationalized in September 2018, just after our study began, as part of a national program to ensure that every upazila had at least one government-run school.

Figure 1: Location of study upazilas in Bangladesh



Notes Map shows unions in Bangladesh defined as "Hard-to-Reach" in terms of water and sanitation coverage (Government of Bangladesh, 2011). Dark grey star indicates Bhola Sadar upazila; light grey star indicates Ramgati upazila.

female. Schools had on average 12.5 teachers, implying a mean student-to-teacher ratio of 46.5, slightly higher than the national average of 34.[8]

Our data confirm that sanitation facilities in the study schools do not meet Bangladeshi or international standards, either in terms of quantity or quality. At baseline, enumerators identified on average 4.5 toilet cubicles available to students per school, of which only 2.9 were functional, equivalent to one for every 226 students. Only three in four schools had any functioning toilet cubicle available exclusively for female students. Fewer than half of schools had any clean, lockable, functioning toilet cubicle available exclusively for female students. No school had any functioning toilet cubicle meeting the minimum guidelines for menstrual health management: exclusively for female students, clean, lockable, and having water, soap, and an adequate container for disposal of menstrual health products inside the cubicle (UNESCO, 2014). Only ten schools had any functioning toilet cubicle that was accessible to disabled students.

Before the intervention, students reported that school sanitation facilities affected female student attendance. Among female students interviewed, only 19% reported that school attendance is unaffected by menstruation: 73% reported that girls leave school earlier when they have their periods, and 8% reported that girls do not usually come to school at all during their period. The majority of girls interviewed (85%) cited poor or inadequate school sanitation facilities among the reasons for which girls miss school when they have their periods, including lack of private facilities for girls to change (64%), to wash cloth pads (51%), and to dispose of sanitary napkins (34%).

What students reported about the consequences of inadequate sanitation facilities contrasts with what school staff reported. When asked about the needs of female students, only 37% of headteachers cited separate toilet cubicles for boys and girls; 23%, a mechanism for the collection and disposal of menstrual health products; and 18%, a space for girls to wash in private. Female teachers were more aware of student needs regarding menstrual health management (MHM) than male teachers.[9] However, women are underrepresented among school staff and management. At baseline, few headteachers (2 out of 60) were female, only one in five school management committee (SMC) members and teachers were female, and ten mixed-gender schools had no female teachers.

3 The AWIS intervention

The Annotated Water Integrity Scan, or AWIS, is an intervention designed by the Water Integrity Network (WIN) and the International Water and Sanitation Centre (IRC) to im-

[8]National data from the UNESCO Institute for Statistics, 2017.

[9]E.g., when asked about the characteristics that a sanitation facility should have for safe MHM, a much higher share of female teachers than male teachers cited the importance of a mechanism for the collection and disposal of sanitary protection materials (56% vs 12%) or of a space to wash privately (62% vs 16%).

prove governance in the water sector (WIN and IRC, 2011). AWIS is designed to measure integrity, defined as a set of practices that impede corruption and promote respect for the rule of law. The practices that characterize integrity fall into three categories or "pillars": transparency, accountability, and participation, as defined in Table 1. AWIS is designed to be replicable and adaptable to water governance at all jurisdictional and geographical scales. Before this study, WIN adapted and piloted AWIS for implementation in several countries and contexts, including rural WASH services in Benin and school WASH services in India. In this study, we apply and evaluate AWIS in the context of high schools in rural Bangladesh, as described in the previous section.

Table 1: AWIS "pillars of integrity"

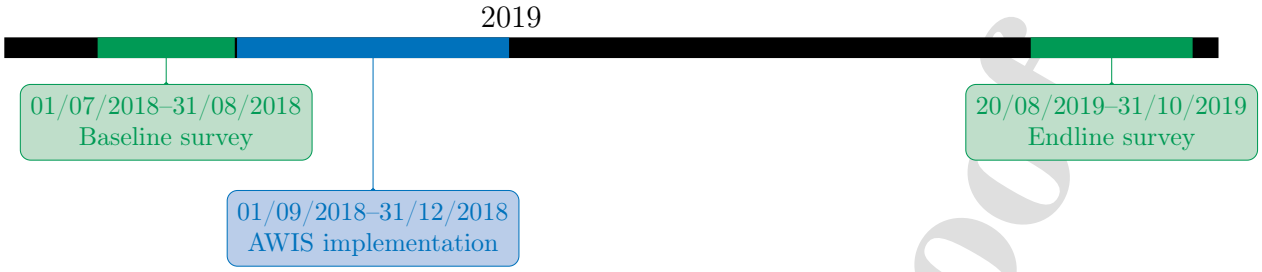
Pillar of integrity	Explanation
Transparency	The existence of written procedures, agreements and contracts that explain the roles and responsibilities of actors.
Accountability	The application of the written procedures and agreements and, where feasible, the potential compliance of actors ("internal accountability").
Participation	The ability of the public, and the users or their representatives (including marginalised and resource-poor groups), to access information, influence decision-making, file complaints effectively and be heard ("external accountability").

Notes "Pillars of integrity" used to characterize "integrity" (from WIN and IRC, 2011).

AWIS is designed to initiate discussions that facilitate the identification of specific and concrete priority steps for improvement. The cornerstone of the intervention is a participatory workshop that brings together stakeholders. The workshop process is designed to facilitate constructive dialogue, to allow discussion of integrity without antagonizing participants with divergent views, to encourage revelation of new information, and to raise awareness of contrasting points of view or unexpected risks. Participating in the AWIS process is intended to help policy-makers and user representatives establish priority actions to enhance water integrity and governance.

The local implementing NGO, Development Organization for the Rural Poor (DORP), carried out the AWIS intervention between September and December 2018 (Figure 2). Local DORP staff facilitated the AWIS workshops and documented the process.

Figure 2: Project timeline



Participants The stakeholders AWIS aimed to bring together comprised students, parents, teachers, the school management committee, social leaders, and local government administrators. By design, four representatives from each of the first four groups should have participated in the workshop. In practice, the number of meeting participants from each group varied slightly (Table 2).[10] In some cases, participants represented multiple groups.

Participating students were drawn from the oldest cohorts, almost all (97%) between 14 and 16 years old. More female students participated than male students, in part reflecting targeting of female-only schools as well as mixed schools. More participants were male than female in all other groups. This gender imbalance partly reflects the gender imbalance among staff and school management committees, but persists even for groups that should be evenly balanced in terms of gender, such as parents. This may reflect prevailing social norms in rural Bangladesh, where female participation in public and political processes is highly constrained (Balk, 1997; Tanjeela, 2021).

The most important deviation from the original project design was that, with one exception, no local government administrators attended any workshops. The exception was one of the first workshops organized by the project, in Bhola, which was attended by a member of the union parishad or council. The absence of local government administrators likely reflects the fragmented nature of the institutional landscape. Government administrators have limited responsibility for improving WASH service provision in schools. Also, many governmental agencies have a limited staff. It may not be surprising that they were unwilling or unable to attend school-level workshops. The DORP facilitators invited social leaders in their place.

[10] In 8 out of 30 cases, the total number of participants was 18 (N = 1), 19 (N = 3), or 21 (N = 4), instead of 20 as per the intervention design.

Table 2: Participants in AWIS workshops

	Mean no. of participants per school		
	Male	Female	Total
Students	1.60	2.47	4.07
Parents	2.37	1.57	3.93
Teachers	2.53	1.10	3.63
SMC members	3.17	0.73	3.90
Social leaders	2.37	1.13	3.50
Headteachers	0.50	0.03	0.53
Local government/administrators	0.00	0.03	0.03
Parents/SMC members	0.10	0.13	0.23
Teachers/SMC members	0.10	0.03	0.13

Notes Mean number of participants in each category participating across schools. One workshop held in each school. SMC: School Management Committee.

Process The AWIS workshop proceeds in three steps:

1. Workshop participants anonymously score the quality of WASH services and WASH-related processes with respect to indicators for transparency, accountability and participation. The scorecard covers five areas: (i) quality of sanitation facilities; (ii) gender; (iii) menstrual hygiene; (iv) facilities for disabled students; and (v) budget and expenditure. For each indicator, local facilitators developed a scale from 1 to 3 where each score is associated with a statement describing quality with respect to that indicator. Participants select which statement best described conditions at their school, yielding an individual score for each indicator between 1 to 3 (Appendix A).
2. Facilitators invite participants to openly discuss the scores for each area, asking participants to share their motivations behind their scores for each item. The objective is not to defend individual opinions but to understand differences in opinions. Based on the outcomes of this discussion, the scores are jointly adjusted and aggregated, until one set of scores is collectively agreed upon.
3. Each participant identifies priority areas for action and specific actions that can be taken within each priority area. Facilitators disclose the results of this exercise to participants and ensure that priority areas and proposed actions are clearly understood by all participants.

Scores Facilitators recorded the scores agreed upon during steps 1 and 2 of the process (Appendix Figure K1). Consistent with the baseline survey evidence described in section 2, workshop participants hold different views about baseline conditions at the study schools. Two randomly-drawn workshop participants from the same school give, on average, the same scores for only half (0.52) of the items, and the within-school range of scores given by partic-

ipants includes both the minimum and maximum score for about half (also, coincidentally, 0.52) of all items. Mean differences across schools explain less than half of the variation in scores across respondents, implying that more than half is driven by variation across workshop participants within schools.[11] While some of this residual variation is likely explained by measurement error,[12] mean scores given by participants correlate positively with direct, independent observations by enumerators during school visits (Appendix Table K1), suggesting that the information captured by the AWIS process is meaningful. Taken together, this descriptive evidence suggests that there is scope for the intervention to improve stakeholder understanding of conditions at the study schools.

Priorities and actions Workshop participants identified, on average, a total of 20 priority areas and 20 specific actions that could be taken.[13] Around half of the most commonly-identified priority areas concerned improvements to physical infrastructure (Appendix Figure K2). The most-commonly identified actions included reaching out to possible sources of funding, awareness-raising, organizing meetings, or delegating responsibility for cleaning and monitoring of WASH facilities to students or teachers. Just under half of schools proposed to allocate a greater share of school funds towards WASH facilities (Appendix Figure K3).

Costs A conservative estimate of the costs of implementing the AWIS intervention is around USD$770 per school, equivalent to USD$1.3 per student. This estimate is conservative because it includes costs such as liaising with local government officials that would likely reduce in proportional terms if the program were implemented at scale. If the intervention successfully increased school attendance, it would have been among the most cost-effective interventions available (J-PAL, 2018).[14] The per school implementation costs are additional to costs of about USD$23,000 for adapting the AWIS intervention for application to schools in rural Bangladesh and about USD$11,500 to train facilitators.

Comparison with other previously-evaluated interventions Tables 3a and 3b compare AWIS to other closely-related "transparency for accountability" and "social accountability" interventions that have been evaluated in other sectors. Like AWIS, these interventions

[11] The (unadjusted) R^2 from a regression of participant scores on school-item fixed effects is 0.46.

[12] Some participants may also not have fully understood all the questions. Also, although scoring is intended to be anonymous, some participants required assistance to complete the scoring process, which might have affected their choices. Note that because scores were recorded anonymously, we cannot evaluate whether participants with different characteristics gave different scores.

[13] Participants listed priorities and actions descriptively. Research assistants categorized responses into discrete categories.

[14] Each school has 575 students with, on average, two more years of secondary education remaining. Had the intervention increased attendance by 2.5 percentage points, a not implausible effect size ex ante, it would have increased total effective schooling by 3.75 years per USD$100 spent, more than any intervention listed by J-PAL (2018) except for deworming in Kenya (Miguel and Kremer, 2004).

typically include elements of *information* and *participation*. The *information* elements can be broadly categorized into three types: 1) general, non-specific information, such as information about rights to public services or responsibilities of service providers; 2) specific information about the service provider (e.g., school or clinic) that is observable to the users or service providers; and 3) specific information about the service provider obtained from external sources, such as surveys or administrative data. The information elements in AWIS belong to types 1 and 2, while type 3 is typically associated with higher costs. The *participation* element comprises some type of meeting or workshop that usually brings service providers and users together. Relative to the other interventions in Tables 3a and 3a, the participatory element of AWIS is also relatively "light touch", comprising a single workshop.

Ex ante strengths of AWIS The AWIS intervention has several features that Kosack and Fung (2014) note are associated with success in transparency interventions. Transparency interventions are more successful: when information is contextualized, for example with information about the rights of service users or responsibilities of the providers defined in national standards; when the information includes subjective impressions as well as objective information; and when the intervention establishes clear actions for users or providers that lead to reform. AWIS is designed to have all of these features: The scoring component of AWIS is defined with respect to Bangladeshi government guidelines with respect to school WASH provision; AWIS includes subjective impressions, such as whether or not girls and their parents feel able to influence the quality of support for MHM; and as noted above, AWIS is designed to establish consensus over specific actions that could be taken to improve outcomes. Thus, ex ante, AWIS appears to have features identified by previous literature as predictive of success.

Table 3a: Interventions that affected development outcomes

Study	Information	Participation	Location	Sector	Time frame
Pandey et al. (2007)	Rights and responsibilities.	NA	India	Health and Social Services	1 year
Pandey et al. (2009)	Rights and responsibilities.	NA	India	Primary Education	2-4 months
Björkman and Svensson (2009), Nyqvist et al. (2017)	Rights and responsibilities, report cards (survey data).	Two community-service provider meetings workshops, held over 1-2 days, 6 months apart, focus on action plan and monitoring.	Uganda	Health	2-4 years
Barr et al. (2012)	Community scorecard.	Scorecard unification process at school-parent-teacher meeting, three termly follow-up meetings	Uganda	Education	Immediately after interventions.
Pradhan et al. (2014)	NA	School management-village council meetings.	Indonesia	Primary Education	About 18 months
Mohanan et al. (2020)	Rights and responsibilities, report cards (survey data).	Monthly community-health committee meetings over 1 year, focus on action plans and adherence to plans.	India	Health	1 year after end of interventions.
Raffler et al. (2020)	Rights and responsibilities, report cards (survey data).	Community-service provider workshop, held over 1 day, 6-monthly followups, focus on action plan.	Uganda	Health	8-20 months
Christensen et al. (2021)	Report cards (administrative data).	Community-service provider workshop, three follow-up meetings, focus on action plan.	Sierra Leone	Health	1.5-4 years.

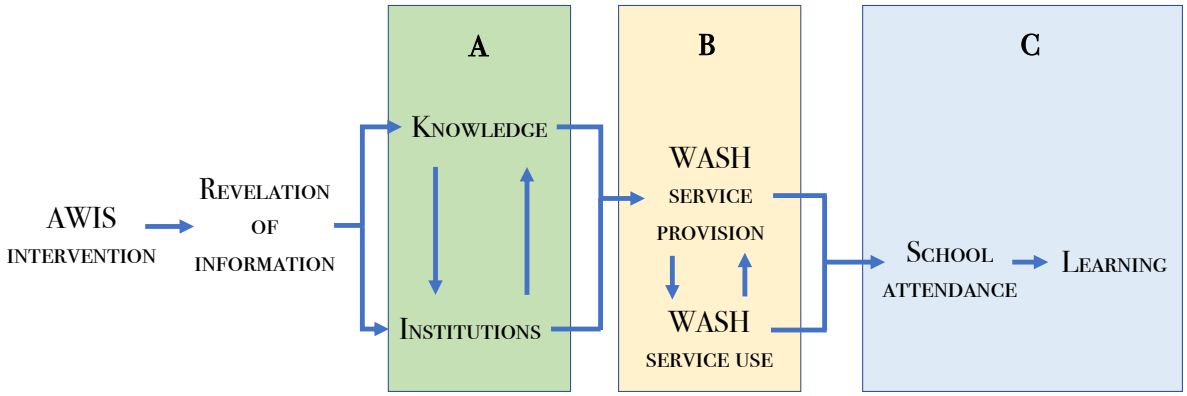
Table 3b: Interventions that did not affect development outcomes

Study	Information	Participation	Location	Sector	Time frame
Olken (2007)	Anonymous comment cards to solicit community views, aggregated by external agents.	Promotion of community-village "accountability" meetings through invitations.	Indonesia	Road construction	Interventions after project, effects through anticipation.
Banerjee et al. (2010)	Rights and responsibilities, report cards (based on community members testing childrens' skills).	Community-headteacher-school committee discussion meeting.	India	Education	1 year
Lieberman et al. (2014)	Report cards for individual children (NGO staff assessments), strategies for parents to get involved locally/nationally.	NA	Kenya	Education	5-6 months
Nyqvist et al. (2017)	NA	Two community-service provider meetings workshops, held over 1-2 days, 6 months apart, focus on action plan and monitoring.	Uganda	Health	2 years
Fabbri et al. (2019)	Report cards (survey data)	Separate meetings with service providers and/or communities.	India	Health	18 months
Arkedis et al. (2021)	Report cards (data gathered by field staff)	6 facilitated community-only meetings, held monthly, focus on action plans.	Indonesia and Tanzania	Health	18 months
BenYishay et al. (2022)	NA	Multi-stakeholder dialogue workshop	Niger	Health and education	10 months

4 Theory of change

Our evaluation design is guided by a theory of change that connects the AWIS intervention to the outcomes of interest (Figure 3). AWIS is designed to spread information about priority areas for action and potential courses of action, and to strengthen institutions. In turn, these changes should improve WASH service provision and may impact on WASH facility use patterns. Improved WASH service provision may in turn increase school attendance and translate into improved exam performance. Drawing on this theory of change, we define three families of outcomes, designated A, B, and C, which are listed in Table 4 and described later in further detail.

Figure 3: Theory of change



The implicit logic underlying transparency interventions is the idea that such interventions can ignite an "action cycle", in which information translates into useful, constructive action that improves service provision (Kosack and Fung, 2014). To initiate such an "action cycle", several criteria must be met. First, the intervention must reveal new information, and the new information must be salient to the users or providers of a service. Second, the revealed information must change the choices or actions of service users or providers. Third, these changes in choices or actions must be sufficient to affect service provision.[15]

To trace out the links between information and action, we provide descriptive evidence about the information that is discussed during the AWIS workshops, actions that are identified during the workshops, and actions that participants report taking after the workshop. Additionally, we evaluate impacts on knowledge about WASH standards and provision, and institutional arrangements (outcome family A). We define institutional arrangements as the presence and quality of institutions for WASH facility management and provision of support to students for MHM. Such institutions include, for instance, whether or not the school has written policies related to WASH issues, whether the school allocates a budget to WASH

[15]We build from a similar discussion in Kosack and Fung (2014).

services, and how frequently school representatives meet with parents. Knowledge and institutions may be self-reinforcing. Improvements to knowledge or institutional arrangements constitute potentially important mechanisms for improvements where we expected that we might detect effects of the intervention even if they are not sufficient to improve service provision over the time frame of our intervention.

We then evaluate impacts on WASH service provision and on use of WASH services (together constituting outcome family B). Understanding effects on service use is essential in order to correctly interpret any effects on WASH service provision because the two are closely related. If WASH service provision improves, use of WASH facilities might increase in response. Increased use might in turn lead to problems with, for example, cleanliness or maintenance that might not have arisen in the absence of changes in use patterns.

The final piece in our theory of change is the link between WASH provision, school attendance, and school attainment (outcome family C). A widespread view is that inadequate WASH provision leads to reduced school attendance, through an increased rate of water-borne disease, and for female students, as a result of students missing school during menstruation (Alam et al., 2017; McMichael, 2019). School attendance is in turn linked to school attainment (e.g., Roby, 2004).

Table 4: Outcomes

	Outcome	Unit of observation	Time fixed effects	Unit fixed effects
A1	Head-teacher knowledge	Headteacher	Upazila-year	School
A2a	Male teacher knowledge	Teacher	Upazila-year	School
A2b	Female teacher knowledge	Teacher	Upazila-year	School
A3a	Male student knowledge	Student	Upazila-year	School
A3b	Female student knowledge	Student	Upazila-year	School
A4	Quality of institutional processes	School	Upazila-year	School
B1a	No. functioning toilets (male student use)	School	Upazila-year	School
B1b	No.functioning toilets (female student use)	School	Upazila-year	School
B1c	No. functioning toilets (mixed use)	School	Upazila-year	School
B2a	Male toilet quality index	Toilet	Upazila-year	School
B2b	Female toilet quality index	Toilet	Upazila-year	School
B3a	Toilet events/day (male student toilets)	Toilet	Upazila-year	School
B3b	Toilet events/day (female student toilets)	Toilet	Upazila-year	School
C1a	School attendance (male students)	Student-day	Grade-upazila-date-year	School-grade-day of week
C1b	School attendance (female students)	Student-day	Grade-upazila-date-year	School-grade-day of week
C2a	Exam results (male students)	Student	Level-upazila-year	School-level
C2b	Exam results (female students)	Student	Level-upazila-year	School-level

Notes Date fixed effects refer to day and month fixed effects. Day of week fixed effects comprise fixed effects for each weekday; schools are closed on Fridays. Year fixed effects refer to calendar year fixed effects, corresponding to periods pre and post intervention.

5 Data

Data collection was managed by a Bangladeshi non-government organization, NGO Forum for Public Health, who were not involved in implementation.[16] NGO Forum for Public Health employed a team of enumerators who collected baseline data before the intervention, during July and August 2018, and endline data between August and October 2019, between 8 and 12 months after the intervention.

One male enumerator and one female enumerator visited each study school for the primary data collection activities, which took one day in each school. We designed the data collection process to minimize disruption to the school day, with a flexible schedule that allowed enumerators to arrange data collection and interviews around student and teacher activities. The data collection process comprises surveys with three main classes of respondents and four data collection modules at both baseline and endline, which we describe briefly below.

Headteachers gave informed consent to the study as a whole and separately to each of the data collection modules. Headteachers in all study schools consented to the data collection process, although in a few cases they declined consent for one or more of the data collection modules.

Enumerators surveyed headteachers, teachers, and students. Female enumerators interviewed female respondents. All interviewees individually assented or consented to be surveyed, although we did not ask respondents about private or sensitive information. In particular, we phrased all questions related to WASH and MHM practices in general terms, asking respondents to draw on both their own and their peers' experiences in their responses.[17]

Enumerators asked all respondents to describe the quality of the existing WASH facilities in the school; the constraints faced by female students during menstruation and the consequences for school attendance; and their views about how to improve WASH services and practices in the school. Additionally, enumerators asked respondents for information about specific areas of interest, as follows:

Headteachers Enumerators interviewed headteachers regarding school programs and initiatives related to WASH and MHM; student, parent and teacher involvement in WASH facility management; and the school budget and institutional processes, especially with respect to WASH facilities.

Teachers Enumerators interviewed male and female teachers in each school regarding the content of school curricula related to WASH and MHM; the system of responsibilities within

[16] See `http://www.ngof.org`.

[17] Appendix B describes the consent process.

the school for cleaning of WASH facilities; and student awareness of WASH and MHM practices. In ten schools, there were no female teachers, so we interviewed only male teachers.

Students Enumerators interviewed male and female students regarding whether and how WASH and MHM practices are discussed and taught in school. To identify students likely to be well-informed and comfortable with an interview, headteachers identified "class captains", a position assigned by teachers to the most responsible and best-performing students. The vast majority of interviewed students were in 10th grade and between 14 and 16 years old.[18]

Enumerators also interviewed school cleaning staff regarding the availability of cleaning products and the cleaning and maintenance schedules. In practice, however, few schools had cleaning staff present on the day of the interview. In particular, at baseline, only 7 schools had any cleaning staff of which which 5 had cleaning staff present on the day of interview.[19]

Enumerators also collected data directly on WASH facilities, WASH use, and student attendance via school records and direct observations.

WASH facility census Enumerators observed and photographed all WASH facilities in study schools, including drinking water points, toilet cubicles, urinals, and handwashing stations. Enumerators recorded whether each facility complied with basic WASH and MHM standards. We also tested all sources of drinking water available in the school for contamination with fecal bacteria and arsenic. Table 5 enumerates the number of school facilities of each type at baseline and endline.

Latrine usage We used event loggers and sensors to monitor latrine usage over three school days in an anonymous, discreet, and non-intrusive way.[20] We tracked latrine usage in a sample of toilet cubicles that were functioning and accessible to students.[21] We primarily used magnetic sensors to record opening and closing of latrine doors. Processing these data allows us to recover a history of latrine use events, corresponding to periods when the door was closed bookended by periods when the door was open to allow entry to and exit from the latrine. Sensors and event loggers are small and were discreetly packaged. Importantly for privacy, the door sensors can be installed outside of the latrine cubicle. Enumerators installed and removed sensors at times when they were not in use by students, in order to interfere as little as possible with normal latrine use.[22]

[18] Two interviewed students were in 9th grade; three were aged 17 years old; and one was aged 13.

[19] We successfully interviewed a member of the cleaning staff whenever one was present.

[20] Specifically, we collect data over three days, excluding holidays and Fridays, when schools are closed.

[21] This implies that we did not collect data if there was no functioning toilet accessible to students.

[22] At baseline, we used motion sensors in some cases when latrines did not have functional doors. Motion sensors use infra-red sensors to detect the presence of a warm body. See Appendix C for further details.

Table 5: Number of observations: sanitation facilities

	Number of observations					
	Baseline			Endline		
Facility type	Male	Female	Mixed	Male	Female	Mixed
Toilet cubicle (in use)	54	95	7	47	103	6
Toilet cubicle (out of use)	-	-	70	-	-	61
Urinal	62	-	-	54	-	-
Hand-washing station	63	86	57	50	74	68
Drinking water point	-	-	60	-	-	60
Total	179	181	194	151	177	195

Notes Table shows number of facilities available for student use enumerated in each category. Toilets that are out of use are listed as for mixed use since gender of users is not defined.

Although we designed this data collection process to minimize intrusion, it is still possible that students might have avoided use of the school latrines during monitoring, a type of "Hawthorne" effect or response to monitoring. However, monitoring was identical in both treated and control schools, and it seems unlikely that students in treated schools would have responded differently to monitoring. As a result, any "Hawthorne" effect should not affect the comparison between treated and control schools.

Attendance records Enumerators scanned and digitized school attendance records for the month in which the survey takes place and additionally for the preceding February and March. We selected these months for comparison across years because no major religious festivals or national exams interrupted school attendance patterns during these months in 2018 or 2019.

Observed attendance To validate the attendance records, enumerators collected direct observations of student attendance on the day of the survey. To observe attendance discreetly, enumerators visited classrooms and asked students to list three improvements they would like to make in their school. Students then voted for their preferred suggestions. Enumerators recorded the number of male and female students voting for each suggestion, allowing them to record the total numbers of male and female students present in class. Enumerators made clear that the voting exercise was purely hypothetical.

Tables 6a and 6b summarize sampling strategies and response rates.[23] Response rates were high, with one exception. The original baseline protocol sampled more teachers than enumerators could feasibly interview on the day of the visit, and the baseline survey form did not enforce compliance with this protocol. Although response rates were low relative to the targeted numbers, enumerators did successfully interview at least one male and at least one female teacher in all study schools with teachers of the given gender. Baseline teacher response rates are similar at baseline in treatment and control schools (Appendix Table K2).[24] We addressed these issues in the endline survey.

We also collected data from three additional sources: administrative data on exam results, workshop records, and semi-structured interviews in a subsample of schools.

Exam results Enumerators downloaded individual exam results for students at study schools from official records publicly available online. Enumerators recorded individual student results (grade point averages, or GPAs) on the most recent standardized national exams for grade 8 (Junior School Certificate, JSC) and grade 10 (Secondary School Certificate, SSC), inferring gender from student names, which are mostly gender-specific in Bangladesh.[25]

Workshop records Staff from DORP, the local implementing NGO, completed a detailed report after each workshop, including participant information, workshop processes, discussion outcomes, and the main lessons learned. DORP staff translated the reports into English. We digitized and classified data from these reports, which we use in descriptive analyses.

Semi-structured interviews To obtain qualitative insights into the strengths and weaknesses of the intervention, enumerators conducted semi-structured phone interviews with up to 4 respondents from a total of 12 schools, evenly sampled at random from treatment and control schools in the two study upazilas. We conducted the interviews in fall 2020, a year after the endline data collection and two years after implementation. Schools were closed for in-person instruction at the time of the semi-structured interviews due to the COVID pandemic. In most schools, we successfully interviewed the headteacher, a parent representative from the school management committee, one male teacher, and one female teacher.[26]

[23] Further details are in Appendix D. Our study design also oversampled AWIS student participants, but in practice this led to only two additional student interviews.

[24] Teachers listed later in the roster were less likely to be successfully interviewed (Appendix Table K3).

[25] Both sets of results are publicly available online on the education board's web-based result publication system. School-level records list individual student results and student ID numbers. Detailed individual student data are also publicly available, including student names, searchable using the student ID number.

[26] We originally planned to conduct in-person focus group discussions in study schools in spring 2020. Because of the COVID pandemic, we initially postponed these activities and then replaced them with phone interviews to minimize unnecessary travel and in-person contact. School closures prevented us from interviewing students. See Appendix E for further details.

Table 6a: Sampling strategies, sample sizes, and response rates: survey modules

Module	Unit of observation	Sampling	Number of observations and response rates				Primary reasons for non-response
			Baseline		Endline		
			Male	Female	Male	Female	
Headteacher survey	Headteacher	Census	58 (100.0%)	2 (100.0%)	58 (100.0%)	2 (100.0%)	N/A
Teacher survey	Teacher	Random sample of one M and one F teacher, plus oversampling of WASH/MHM- responsible teachers and AWIS participants.	124 (66.0%)	52 (55.3%)	136 (91.3%)	76 (100.0%)	Baseline coding/ protocol error, teacher non-consent.
Student survey	Student	Random sample of one M and one F class captain from the highest grade with a class captain present during the survey.	48 (98.0%)	60 (100.0%)	49 (100.0%)	60 (100.0%)	Student non-assent (1 student).

Notes M: male, F: female. All numbers of observations correspond to numbers of successfully completed surveys. Response rates defined relative to number of respondents targeted for interview and shown in brackets. Primary reasons for non-response list the most commonly-occurring reasons.

Table 6b: Sampling strategies, sample sizes, and response rates: direct observation modules

Module	Unit of observation	Sampling	Number of observations and response rates				Primary reasons for non-response
			Baseline		Endline		
			Male	Female	Male	Female	
WASH facilities	WASH facility	Census of toilets, urinals, handwashing facilities, and water points.	554 facilities in 60 schools (100.0%)		523 facilities in 60 schools (100.0%)		N/A
Latrine usage	Toilet	Up to 4 toilets that are functioning and accessible to students; randomly sampled if more than 4.	46 toilets (85.2%)	82 toilets (90.1%)	32 toilets (74.4%)	87 toilets (90.6%)	Enumerator or sensor error.
Attendance records	Student	One class per grade, randomly sampling between classes if more than one class.	8532 students in 234 classes (95.5%)	14085 students in 289 classes (96.3%)	9791 students in 244 classes (99.6%)	15183 students in 299 classes (99.7%)	Headteacher non-consent (2 schools).
Attendance observations	Student	One class per grade, randomly selecting between classes if more than one class.	3955 students in 234 classes (95.5%)	7720 students in 289 classes (96.3%)	4291 students in 235 classes (95.9%)	7968 students in 290 classes (96.7%)	Headteacher non-consent (2 schools).
Exam results	Student	Administrative data	5181 students in 49 schools (100.0%)	5666 students in 60 schools (100.0%)	5039 students in 49 schools (100.0%)	6002 students in 60 schools (100.0%)	N/A

Notes Number of observations for attendance records corresponds to number of students for whom we have attendance data in February and/or March of the corresponding year. Response rates shown in brackets. Response rates for WASH facilities and exam results calculated at the school level. Response rates for latrine usage relative to the toilets selected for data collection according to the protocol; mixed gender toilets (7) pooled with male-only toilets for the purposes of reporting sample size and response rates in this table. Response rates for attendance observations are at the class level.

6 The experiment

We randomly assigned 30 schools to receive the intervention and 30 schools to a control group that received no intervention via this project, although importantly we did not prevent them from receiving any other interventions. All schools had the same ex ante probability of receiving the program.

We assigned schools to treatment and control at public lotteries, ensuring transparency over the treatment assignment process. We held one treatment assignment lottery in each of the two study upazilas, Bhola Sadar and Ramgati, with 15 schools in each upazila assigned to treatment and 15 to control. The lotteries took place after baseline data collection.

Figures 4a and 4b shows the location of treated and control schools in Bhola Sadar and Ramgati, respectively. Balance tests using baseline data confirm that differences between schools assigned to treatment and control are consistent with differences that could arise due to chance. Table 7 shows that the treatment and control groups are balanced with respect to the main outcomes at baseline.[27]

A potential threat to experimental validity arose because of a second intervention that was carried out in schools during this time period, consisting of distribution of free sanitary products to schools for female students and education about menstrual hygiene. This intervention was carried out in 5 study schools before baseline and 13 study schools between baseline and endline. We successfully coordinated with the implementing organization to ensure that assignment to this intervention was uncorrelated with treatment with the AWIS intervention (Appendix Table K11).

[27]To report tests of joint significance of differences across outcomes, we collapse all data to weighted school-level means, using sampling weights as we do throughout the analysis. To avoid missing observations, we pool data from male and female respondents, and from toilet facilities used by male, female, and mixed students. When we pool data across genders, we use indices constructed on pooled data. We residualize all outcomes with respect to upazila fixed effects, interacted with gender, grade, exam level, and date fixed effects where relevant, for comparability with the randomization checks on individual outcomes. For latrine use data, we include sensor-type fixed effects as we do for the individual checks. We replace missing data with the upazila-level mean value: for one school for outcome B2, five schools for outcome B3, and one school for outcome C1. Appendix Tables K4 to K10 provide extended baseline comparisons between schools, including additional joint tests of comparability on 20 school-level characteristics.

Figure 4a: Study schools in Bhola Sadar upazila

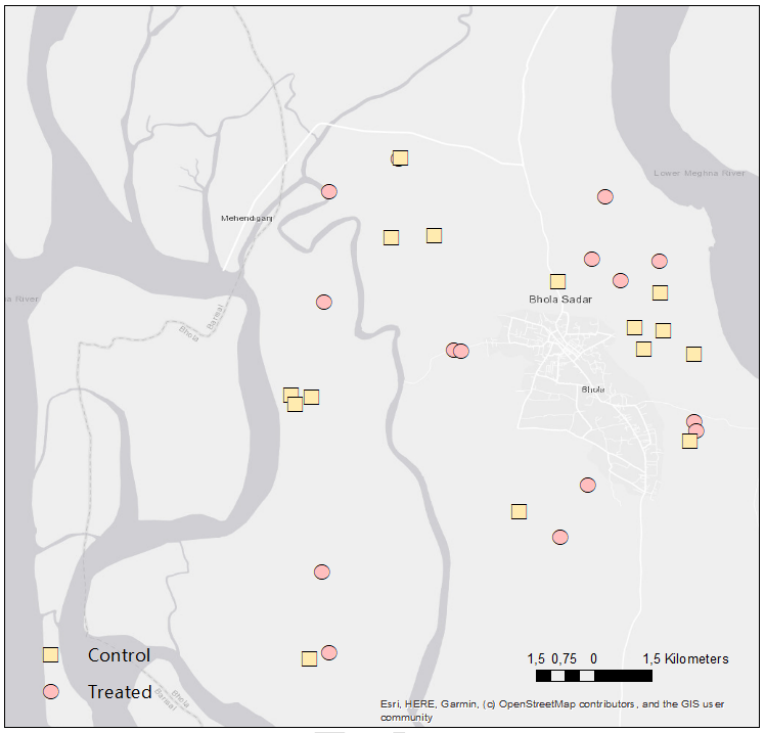


Figure 4b: Study schools in Ramgati upazila

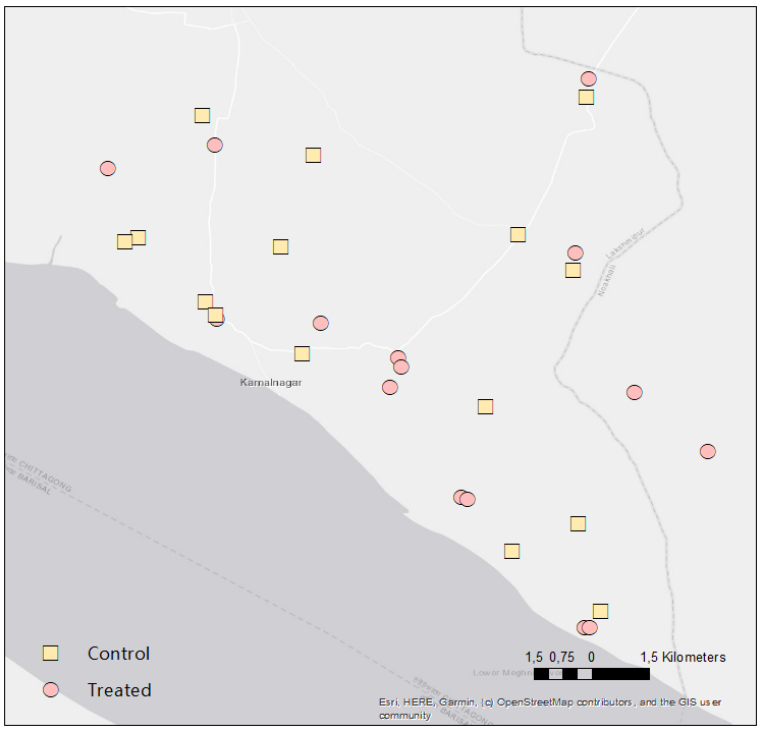


Table 7: Baseline summary statistics

	Control	Treated	Obs
A1 Head-teacher knowledge	0.000	-0.034	60
	(0.113)	(0.100)	
A2a Male teacher knowledge	-0.000	-0.094	124
	(0.088)	(0.099)	
A2b Female teacher knowledge	-0.001	0.378*	52
	(0.127)	(0.133)	
A3a Male student knowledge	-0.013	-0.073	48
	(0.087)	(0.142)	
A3b Female student knowledge	0.000	-0.098	60
	(0.106)	(0.121)	
A4 Institutional quality	-0.003	0.006	60
	(0.066)	(0.061)	
B1a No. functioning toilets (male student use)	1.200	1.008	49
	(0.256)	(0.199)	
B1b No. functioning toilets (female student use)	1.267	1.900**	60
	(0.165)	(0.165)	
B1c No. functioning toilets (mixed use)	0.200	0.033	60
	(0.082)	(0.037)	
B2a Male toilet quality index	-0.011	-0.044	92
	(0.059)	(0.075)	
B2b Female toilet quality index	-0.004	0.063	127
	(0.054)	(0.056)	
B3a Toilet events/day (male student toilets)	10.113	9.000	80
	(2.861)	(1.760)	
B3b Toilet events/day (female student toilets)	11.603	14.312	114
	(2.735)	(3.195)	
C1a Daily attendance (male students)	0.602	0.554*	320988
	(0.020)	(0.021)	
C1b Daily attendance (female students)	0.673	0.644	552949
	(0.017)	(0.015)	
C2a Exam result (male students)	2.171	2.385*	5131
	(0.083)	(0.079)	
C2b Exam result (female students)	2.296	2.251	5586
	(0.079)	(0.039)	
p value of F-test for joint significance		0.456	60
p value of Hotelling's T-Squared test		0.455	60

Notes Standard errors clustered by school and shown in parentheses. Results from regressions with stratification controls. Unit of observation as listed in Table 4. All observations weighted so that each school (or school and grade/level, for outcomes in family C) counts equally in summary statistics. Attendance comparisons include day-grade-upazila fixed effects; exam result comparisons include exam level-upazila fixed effects; all other regressions include upazila fixed effects. Latrine use comparisons also include sensor type fixed effects. p values from joint F test and Hotelling's T-squared test joint significance of differences on all outcome variables between treatment and control groups. *** p <0.01, ** p <0.05, * p <0.1.

7 Empirical approach

Estimating equation We analyze data as a panel or a repeated cross-section. When the unit of observation is the school, the data form a balanced panel. When the unit of observation is more granular, such as at the student level, the data form a repeated cross-section, because both the population and the sample may change between baseline and endline, for example because some students graduate and new students enroll. We estimate the following equation:

$$Y_{ijt} = \alpha_j + \gamma_t + \beta \left(\mathbb{1}_{treated,j} \times POST_t \right) + \beta_u \left(U_j \times \mathbb{1}_{treated,j} \times POST_t \right) + \epsilon_{ijt} \tag{1}$$

where Y_{ijt} is an outcome variable in unit of observation i in school j at time t. The α_j term is a unit fixed effect that absorbs baseline differences across schools, and γ_t is a time fixed effect that absorbs overall time trends. In the simplest specifications, unit fixed effects are school fixed effects, and time fixed effects are upazila-year fixed effects. When we have more granular data available we additionally account for variation across subgroups or more comprehensively across time (see Table 4). The indicator variable $\mathbb{1}_{treated,j}$ takes the value one if school j is assigned to treatment and zero otherwise, the indicator variable $POST_t$ takes the value one at endline and zero at baseline, and the variable U_j is a demeaned upazila dummy that accounts for stratification in random assignment to treatment. The coefficient of interest is β, the difference-in-differences between treated and control schools. Including the centered and interacted stratification control ensures that β is estimated consistently (Gibbons et al., 2019; Imbens and Rubin, 2015; Lin, 2013). Random assignment of treatment to communities ensures that $\mathbb{1}_{treated,j}$ is uncorrelated with the error term ϵ_{ijt}, implying that estimates of β have a causal interpretation.

Hypothesis testing The number of schools in our study is small, so we use randomization-based inference to evaluate the statistical significance of the estimated coefficient, $\hat{\beta}$. We simulate the lottery that assigns schools to treatment and control 1000 times. We then estimate the distribution of $\hat{\beta}$ under the sharp null hypothesis of no difference between treated and control schools that is imposed by the simulations. We then compare the realized differences between treated and control schools to the distribution of differences under the null. Because the number of schools in our study is relatively small and the direction of the main hypotheses was clear—the intervention is designed to improve outcomes—we pre-specified one-sided hypothesis tests to maximize power to detect improvements. The one-sided hypothesis tests evaluate whether or not we can reject the null hypothesis that treatment did not improve outcomes. We also report analytical standard errors, accounting for clustering at the school level.

Weighting The number of observations per school differs for different outcome variables. To ensure consistency between analyses, we always weight observations by the inverse of the number of observations per school.[28] This approach ensures that each school counts equally in summary statistics in all analyses.

Pre-specification We pre-specified our analyses, and we follow a pre-analysis plan.[29] We deviate from this pre-analysis plan in one significant respect. We originally pre-specified that our primary hypothesis tests would pool data from male and female respondents, or toilet facilities used by male and female students, and that we would report effects disaggregated by gender as secondary hypotheses. However, male and female respondents are very different, as are toilets used by male and female students. Female respondents are much more knowledgeable about WASH/MHM, and toilets used by female students have higher quality at baseline (Appendix Figures K4a to K4c). Not all schools have male students and not all schools have female teachers, implying that pooling across genders also involves pooling across different populations. Ex post, we prefer the cleaner and more transparent approach of reporting results separately for each gender.[30] The smaller sample sizes for female teachers and male students are reflected in larger confidence intervals for effects in these groups.

Multiple hypothesis testing We report results for a total of seventeen outcomes in three families. There is thus a relatively high likelihood that we reject the null hypothesis in at least one of these tests due to chance, even if the null hypothesis of no improvement in outcomes is in fact correct. Along with the "naïve" p values, which would be correct for tests of any one of the outcomes in isolation, we also report p values that control the family wise error rate (FWER) within each of our three outcome "families". The FWER-p values we report are obtained from simulating the joint distribution of p values for all outcomes within one

[28] For outcomes A2a and A2b, the weights we apply account for differential probabilities of inclusion in the sample for different types of teacher and specifically for oversampling of teachers with special responsibility for WASH or MHM. For outcome B3a and B3b, weights account for differential probabilities of inclusion in the sample for functioning and non-functioning toilets. For outcomes in family C, we weight observations so that each grade (C1a and C1b) or exam type (C2a and C2b) contributes equally, ensuring that the estimated effects are insensitive to changes in the distribution of students across grades or exam types.

[29] We pre-registered hypotheses on July 17th, 2018, during baseline data collection and before implementation took place. We finalized a pre-analysis plan using the baseline data on September 25, 2019, after endline data collection had begun but before analyzing any endline data. RCT ID: AEARCTR-0003111.

[30] Our original approach also created inconsistencies between results in the pooled samples and the gender-separated samples. Weighting schools equally in the gender-separated samples implies weighting each school and gender equally, whereas weighting schools equally in the pooled regression implies that changes in the gender characteristics of the sample are reflected in the estimated changes in outcome variables. In principle, we could have anticipated this issue when we wrote the pre-analysis plan. In practice, the problem did not occur to us until we implemented the analyses after endline. We also deviate from the pre-specified approach in a second, more minor respect. We replace school-grade-date fixed effects with school-grade-day of week fixed effects for outcomes C1a and C1b. We make this change because our original approach effectively dropped data from dates in 2019 that were Fridays in 2018 and vice versa, about one in six observations. Appendix Table K12 shows the originally pre-specified primary hypotheses and regression specifications.

family and comparing the "naïve" p value for each outcome to the minimum p value across all outcomes in the family (Anderson, 2008; Westfall and Young, 1993).[31] A drawback of disaggregating all results by gender is that it expands the number of hypotheses we analyze, increasing the risk of erroneously rejecting a null hypothesis that is in fact correct. On the other hand, expanding the number of hypotheses we test makes the already-conservative FWER-adjusted results still more conservative.

Joint hypothesis tests Often, the strongest evidence for impacts comes not from large effects for one outcome but from modest but consistent effects across many outcomes (Young, 2019). We thus also report a joint test for each family of hypotheses which ranks the p values from smallest to largest and compares the distribution of each of these to the distribution of ranked p values obtained from randomization-based inference. This allows us to infer the likelihood of seeing differences at least as large, in relative terms, as the observed differences across all outcomes within each family. We did not pre-specify these joint hypothesis tests.

Exploratory analyses Given that we unexpectedly observed cases in which the intervention worsened outcomes relative to the control group, we also report p values from two-sided hypothesis tests. These p values evaluate the likelihood that differences with the realized absolute magnitude would have occurred under the null hypothesis of no effect. These tests are not pre-specified, and our experiment was not powered to detect negative effects, so these results should be treated as exploratory.

Index construction Throughout, we follow Anderson (2008) and use inverse covariance weighting to combine several sub-indices in summary indices. This approach weights information from several standardized outcomes by the inverse of their covariance matrix in order to maximize the amount of information captured in the index. Following the original formulation in Anderson (2008), we calculate the index across all non-missing data for each observation separately, ensuring that we retain the maximum possible amount of information in construction of the index.[32] In all cases, we normalize the index relative to control schools at baseline, so that changes with respect to baseline and treatment effects are measured with respect to pre-intervention conditions. Note that index variables are thus close to zero—or exactly zero when there are no missing data—for control schools in Table 7 by construction.

Power Although the number of treated and control units is relatively small, constrained by the available budget for both the intervention and data collection, we maximize our

[31] Appendix F provides more details on how we implement multiple and joint hypothesis testing. We note that we pre-specified that we would also report p values that control the FWER across the full set of outcomes, but since none of the results are independently statistically significant when we control the FWER within each of the three families, the more conservative p values do not provide much additional information.

[32] We adapt code generously shared by Samii (2018) so that it deals with missing observations.

ability to detect effects by collecting data wherever possible at a more disaggregated unit of observation (e.g., for individual WASH facilities or latrine cubicles or individual students) or over a greater period of time (e.g., over multiple days). Ex post power calculations, using either analytical standard errors or the distribution of the point estimates obtained from randomization-based inference, suggest that the study was powered to detect effects of reasonable magnitude compared to ex ante expectations, typically an increase of 0.2 to 0.4 standard deviations for the outcome variables that are indexes, characterized as small to medium-sized effects (Cohen, 1988).[33]

Robustness tests We implement a series of pre-specified robustness tests. We re-estimate all the main effects using control variables selected by the Lasso algorithm from a pre-specified group of exogenous control variables comprising baseline school-level characteristics interacted with an endline dummy. The results remain very stable in these alternative specifications.[34] In the interests of brevity, we summarize the results of all other robustness tests when we discuss the relevant results.[35]

8 Results

Motivated by the theory of change described in section 4, we evaluate the effect of the AWIS intervention on the following three outcome families: knowledge, processes and institutions (A); WASH provision and use (B); and educational outcomes (C). In the interests of readability, we integrate discussion of how we construct each specific outcome within each family with the discussion of results.

A: Knowledge and institutions

We estimate effects on six outcomes in this family. We measure knowledge of WASH and MHM standards for headteachers (A1), male and female teachers (A2a and A2b), and male and female students (A3a and A3b) using indices. Each index is created from several sub-indices that measure the share of correct responses given to a number of questions about best practices for WASH and MHM. For teachers and students, we construct indices separately

[33]See Appendix Table K13. We calculate minimum detectable effects for one-sided hypothesis tests at 90% significance and 80% power as 2.12 times the estimated standard error of the point estimate (Djimeu and Houndolo, 2016), using both the analytical standard errors and the distribution of point estimates from the randomization-based inference; both approaches give similar conclusions. Note that while ex post power calculations that use estimated effect sizes are known to be misleading (see, e.g., Gelman, 2019), this problem does not affect ex post power calculations that use only information about standard errors.

[34]See Appendix G for details of how we implemented this approach and results tables estimated using this approach.

[35]We also implement robustness tests that we did not prespecify. When we do so, we note that this is the case.

for male and female respondents. We measure institutional quality using an index capturing the presence and quality of institutions to manage WASH facilities and provide support to students for MHM (A4).[36]

Results are shown in Table 8. Most (5 out of 6) of the indices show small to moderate positive treatment effects. The improvements in knowledge are largely driven by changes in knowledge about MHM while the improvements in institutions are driven by increased frequency of parent meetings and an increase in the likelihood that the school has a budget specifically assigned to cleaning WASH facilities.[37] Two of the individual treatment effects are statistically significant at the 10% level, although neither of these effects are statistically significant when we control the family-wise error rate (FWER). The likelihood of seeing effects of at least this magnitude across all outcomes is relatively low, however. Under the sharp null of no positive effects on any outcome, the probability of seeing this distribution of p values is .031 (see also Appendix Figure K5). Thus the results seem most consistent with modest positive effects on knowledge and institutions across a range of outcomes.

These results are also consistent with our descriptive data. Semi-structured interview respondents and field staff from DORP (the implementing NGO) reported that some schools painted the government guidelines for school WASH on school walls after the AWIS workshop. Participants also reported that the workshop helped them understand their rights and responsibilities with respect to WASH (Appendix Figure K6).

The main sample for male and female teachers comprises teachers selected at random and teachers selected because they have special responsibility for WASH or MHM education. These groups are sampled consistently across treatment and control groups. Just over two-fifths of the teachers in the main analysis in treated schools themselves participated in the AWIS workshop. Improvements in teacher knowledge are driven by the randomly-sampled teachers rather than those with special responsibility for WASH or MHM (Appendix Table K18). The results are not sensitive to excluding new staff from the sample (Appendix Table K19).[38]

[36]Detailed information on the construction of all outcomes in family A is provided in Appendix H.
[37]Appendix Tables K14 to K17 report effects broken down by subindex.
[38]We did not pre-specify the robustness tests desribed in this paragraph.

Table 8: A: Knowledge and institutions

	Head-teacher	Teacher knowledge		Student knowledge		Institutional
	knowledge	Male	Female	Male	Female	quality
Treatment × post	0.264	0.121	-0.032	0.123	0.249	0.109
	(0.169)	(0.164)	(0.223)	(0.213)	(0.177)	(0.132)
One-sided p value	0.075	0.235	0.560	0.275	0.085	0.213
FWER one-sided p value	0.371	0.628	0.628	0.628	0.371	0.628
Two-sided p value	0.133	0.466	0.889	0.553	0.166	0.430
N	120	226	111	96	120	120
Baseline mean	-0.008	-0.041	0.179	-0.062	-0.047	0.000
Change in control schools	-0.191	-0.104	0.175	-0.040	0.080	0.178
Change in treated schools	0.072	0.018	0.143	0.093	0.330	0.287
Unit of observation	Headteacher	Teacher	Teacher	Student	Student	School

Notes Standard errors clustered by school and shown in parentheses. Results from regressions that include school and upazila-year fixed effects and stratification controls. Weights applied so that all schools have the same weight in summary statistics. One-sided p values (pre-specified) test the null hypothesis of no positive effects; FWER-controlled p values (pre-specified) hold the possibility of making type I errors within the family to its nominal level. Two-sided p values (exploratory) test the null hypothesis of no effects of any sign. Changes in treated and control schools are raw weighted mean changes between baseline and endline for each group, and the difference between the two may not correspond to the estimated treatment effect.

B: WASH service provision and use

We measure effects on seven outcomes within this family (Table 9). First, we enumerate the number of functioning toilets that are available to students in each school, disaggregating these results by the number of toilet cubicles available to male students only (B1a), female students only (B1b), and for mixed use (B1c). A toilet cubicle is an individual stall, seat, squat-plate or drop-hole where a person can defecate in private. Toilets are defined as functioning if they can be used and are not broken, damaged, or full. Relative to the control group, we cannot reject the null hypothesis of no positive impact on the number of functioning toilets available to any group. The number of functioning toilets for female student use actually declines in relative terms, primarily driven by improvements in the control schools rather than absolute declines in the treated schools.

We then examine the quality of toilets available to students. We create an index which measures the average quality of WASH facilities, including functionality, cleanliness, and, for toilets used by female students, the presence of items required for hygienic practices for MHM.[39] Since few schools have mixed gender toilets, we report this index only for toilets available to male students (B2a) and female students (B2b). We cannot reject the null of no quality improvement for either group, and quality declines for toilets available for female students. Again, these effects are primarily driven by relatively large improvements in quality in control schools, although toilet cleanliness declines in absolute terms in treated schools.

We finally turn to use of latrine facilities. Increased use of latrine facilities might be a proxy for improved quality, with the complicating factor that increased use might itself decrease cleanliness. As prespecified, we measure latrine use as the count of toilet use events in each toilet, coding non-functioning toilets as zero use, in toilets available to male students (B3a) and female students (B3b). Again, we cannot reject the null hypothesis of no increase for either group, and usage declines relative to the control group in toilets available to female students. The point estimate for girls remains negative across a range of alternative approaches to processing the sensor data.[40]

Our study was designed and powered to detect effects of the expected (positive) sign. We thus pre-specified one-sided hypothesis tests. However, all three of the treatment effects for toilets available to female students are negative, and two would have been rejected at the 10% level had we pre-specified two-sided hypothesis tests. The likelihood of seeing three negative effects of this magnitude within this outcome family and gender is small ($p = .008$, Appendix Figure K7). In all three cases, the negative point estimate is driven not by absolute declines in the treated group but by relatively large improvements in the control group.

[39] The full set of sub-indexes is in Appendix I. Sub-index results are in Appendix Tables K20 and K21.

[40] Appendix Table K22 describes sensitivity to excluding non-functional toilets (not pre-specified) and controlling for the type of sensor used to detect use at baseline. Appendix Table K23 shows results using an alternative measure of use, total duration of use events. Appendix Table K24 shows results aggregating latrine use frequency or duration across latrines to the school level.

Table 9: B: WASH provision and use

	Functional toilets			Index WASH quality		Toilet event frequency	
	Male	Female	Mixed	Male	Female	Male	Female
Treatment × post	0.228	-0.533	0.067	0.107	-0.163	-0.112	-8.322
	(0.301)	(0.284)	(0.178)	(0.147)	(0.087)	(4.516)	(6.202)
One-sided p value	0.252	0.950	0.454	0.244	0.965	0.496	0.927
FWER one-sided $ value	0.889	1.000	0.967	0.889	1.000	0.967	1.000
Two-sided p value	0.461	0.083	0.876	0.492	0.082	0.987	0.175
N	98	120	120	162	238	128	205
Baseline mean	1.079	1.583	0.117	-0.007	0.045	5.821	11.805
Change in control schools	-0.292	0.400	-0.033	-0.049	0.157	1.736	14.835
Change in treated schools	-0.040	-0.133	0.033	0.020	-0.043	-0.523	8.750
Unit of observation	School	School	School	Toilet	Toilet	Toilet	Toilet

Notes Standard errors clustered by school and shown in parentheses. Results from regressions that include school and upazila-year fixed effects and stratification controls. Weights applied so that all schools have the same weight in summary statistics. One-sided p values (pre-specified) test the null hypothesis of no positive effects; FWER-controlled p values (pre-specified) hold the possibility of making type I errors within the family to its nominal level. Two-sided p values (exploratory) test the null hypothesis of no effects of any sign. Changes in treated and control schools are raw weighted mean changes between baseline and endline for each group, and the difference between the two may not correspond to the estimated treatment effect.

C: School attendance and educational attainment

We finally turn to effects on school attendance and educational achievement (Table 10). Our primary measure of attendance is drawn from school records from the months of February and March, two months during which school attendance was relatively uninterrupted by exams or national holidays in 2018 and 2019. We estimate effects on school attendance for male (C1a) and female (C1b) students. In both cases, attendance rates in treated schools increase by a small amount (0.9 percentage points, or p.p.) relative to control schools, but in neither case can we statistically reject the null hypothesis of no improvement.[41] The 90% confidence intervals exclude increases of more than 4.1 p.p for girls. For comparison, data on self-reported absence during menstruation suggest that absence associated with menstruation could reduce female student attendance by 4.3 p.p.,[42] while the extent of absence associated with water-borne disease may be larger but is more uncertain (McMichael, 2019).

One might be concerned that registered attendance is inaccurate or prone to overreporting, especially as female student attendance is incentivized through a stipend scheme.[43] However, observed and registered attendance are similarly distributed and highly correlated (Appendix Figures K8a and K8b). At baseline, overreporting is absent for female students and modest for male students. In schools and grades consisting only of one class—in which we can be sure that both measures correspond to the same group of students—around 1.7 additional male students are recorded present per class than enumerators observe (Appendix Table K26). Overreporting increases between baseline and endline but does not increase differentially in treated and control schools (Appendix Table K27).

Our primary measure of educational achievement is GPA on national standardized tests. We compare results from standardized tests—national Junior and Senior School Certificate exams—taken before and after the intervention. In both cases, we fail to detect increases in test results for either male (C2a) or female (C2b) students. The 90% confidence intervals rule out improvements of 0.16 points on a 5-point scale for female students and 0.07 points for male students. We also do not detect effects on the share of students passing the exam, an alternative measure of achievement (Appendix Table K28).[44]

[41] The results are similar, but make less efficient use of the data, when we use the pre-specified school-grade-date fixed effects instead of school-grade-day of week fixed effects (Appendix Table K25).

[42] Alam et al. (2017) report that 41% of Bangladeshi school girls report missing school due to menstruation and that those who report missing school miss on average 2.8 days per menstrual cycle. Assuming a median cycle length of around 31 days for adolescent girls (Flug et al., 1984) and a 6 day school week, this is equivalent to a 4.3 percentage point reduction in attendance.

[43] The Female Secondary School Stipend Project pays tuition fees and provides monthly stipends for unmarried rural girls up to grade 10 who attend recognized institutions, remain unmarried, and meet attendance and exam result targets.

[44] Students fail the exam if they score a mean GPA of below 1. For these students, the final score is not recorded. In the main analysis, we code this students as having a GPA equal to 0.5, but the results are insensitive to alternative assumptions (Appendix Table K29).

Table 10: C: School attendance and exam results

	School attendance		Exam results	
	Male	Female	Male	Female
Treatment × post	0.010	0.006	-0.064	0.021
	(0.026)	(0.021)	(0.079)	(0.083)
One-sided p value	0.367	0.431	0.774	0.411
FWER one-sided p value	0.776	0.776	0.776	0.776
Two-sided p value	0.714	0.795	0.442	0.827
N	695623	1152921	10119	11471
Baseline mean	0.574	0.658	2.322	2.278
Change in control schools	0.017	0.005	0.168	0.146
Change in treated schools	0.037	0.012	0.071	0.216
Unit of observation	Student-day	Student-day	Student	Student
Time fixed effects	Grade-upazila-date-year	Grade-upazila-date-year	Level-upazila-year	Level-upazila-year
Unit fixed effects	School-grade-day of week	School-grade-day of week	School-level	School-level

Notes Standard errors clustered by school and shown in parentheses. Results from regressions that include school and time and unit fixed effects as specified and stratification controls. Weights applied so that schools and grades (for school attendance) and schools and exam levels (for exam results) have the same weight in summary statistics. One-sided p values (pre-specified) test the null hypothesis of no positive effects; FWER-controlled p values (pre-specified) hold the possibility of making type I errors within the family to its nominal level. Two-sided p values (exploratory) test the null hypothesis of no effects of any sign. Changes in treated and control schools are raw weighted mean changes between baseline and endline for each group, and the difference between the two may not correspond to the estimated treatment effect.

A caveat to these results is that students not enrolled in the school or grade can sit the School Certificate exams. This might in principle attenuate any treatment effects of the program, because some students sitting the exams will not have been exposed to treatment. Also, not all enrolled students sit the exam. However, the number of students sitting the exams prior to the baseline survey is highly correlated with current class size in the relevant grade, suggesting that any attenuation may be small (Appendix Figures K9a and K9b).

The results in outcome family C could be affected if the intervention led treated schools to increase or decrease enrolment relative to the control group. Male student enrolment increases modestly and by similar amounts in both treatment and control schools. Female student enrolment actually declines in treated schools and increases in control schools (Appendix Table K30). The total number of students sitting exams is relatively stable over time and does not change differentially in treated and control schools (Appendix Table K31).

9 Discussion

The results show that the AWIS intervention led to modest improvements in knowledge and institutions for governing and managing WASH services but did not improve WASH service provision. Exploratory analysis suggests that treated schools may have improved WASH service provision for female students less successfully than control schools. In this section, we suggest four potential reasons why the intervention did not improve WASH service provision despite the positive effects on knowledge and institutions. One reason relates to the research design, and three relate to the intervention design. We draw upon three sources of descriptive data: 1) participant evaluations collected during the AWIS workshop; 2) retrospective survey questions about the workshop and participant behaviour afterward; and 3) qualitative data from the semi-structured interviews. We also compare our results with previous literature to shed light on the generalizability of our findings.

Time frame of the evaluation We evaluate the effects of the intervention in the year following the workshop. This time frame is not atypical compared to previous studies (Tables 3a and 3b). However, some potential effects may continue to evolve over longer time frames, such as the construction of new WASH facilities or the incorporation of improved WASH facilities in new construction. Our results leave open the possibility that the improvements we see in knowledge and institutions might still lead to meaningful impacts on WASH provision over longer time frames.

We find some evidence to support this view in our descriptive data. The intervention increases the similarity of answers given by different respondents to questions about the problems faced by female students related to menstrual hygiene management (MHM), suggesting

an improved common understanding of problems with WASH provision at study schools.[45] Almost all workshop participants we surveyed at endline also reported taking action after the workshop, and almost all felt that their actions were successful (Appendix Figure K10). In one study school, semi-structured interview respondents consistently describe transformational changes in WASH provision, including physical improvements to infrastructure and wide-ranging improvements in collective facility management. These changes appear to have taken place after the endline survey. Interview respondents attribute the impetus for these changes to the AWIS intervention. This pattern of results is also consistent with findings after a similar intervention conducted in Indonesia and Tanzania: focus groups in a substantial minority of communities described sustained, concrete efforts two years later, despite the intervention doing little on average to improve service provision in measurable terms (Arkedis et al., 2021; Kosack et al., 2019).

On the other hand, these self-reported data could reflect experimenter demand effects or other types of reporting bias. While the implementing NGO was not involved in data collection, the NGO that collected data might have been associated with the intervention through their role in organizing the treatment assignment lotteries. The dichotomy between the self-reported and measured efficacy of actions taken by participants could also arise through attribution bias if participants mistakenly attribute unrelated progress to the effects of their actions (Kosack et al., 2019). Other evidence also cautions against too much optimism. Some potential improvements, such as changes to cleaning protocols, are quite low-cost in terms of budget, albeit not necessarily in terms of time or effort, but few schools appear to have successfully implemented similar improvements. Interviewees speak of limited financial resources and the difficulty of raising additional funds when students' families are poor. Facilities also fall rapidly into disrepair. One semi-structured interview respondent speaks of new facilities falling into disrepair in less than eighteen months. The number of schools with no functioning toilets available to students is actually higher at endline (6) than at baseline (2), in both cases evenly divided between treated and control schools. Another notes the challenges of maintaining knowledge and informal institutions in the context of a rapidly changing school population.

Time-limited nature of the intervention One potential reason why the intervention we study might not improve WASH service provision is that it is limited to a single, one-off workshop. Such an intervention design would be attractive, if successful, as cost-effectiveness would then be high. However, the "action cycle" of Kosack and Fung (2014) implicitly requires new information about the consequences of actions to be fed back to stakeholders. Otherwise,

[45] We calculate a similarity index across dyads of survey respondents including headteachers, male teachers, and female students. These are the classes of respondents that we interview in all study schools. For each dyad, we calculate a simple matching coefficient that measures similarity of responses across groups of questions. We estimate an increase in dyad-level similarity in responses to questions about how MHM impacts female students. More details are in Appendix J.

stakeholders will not know whether or not their actions were effective, and efforts may not be sustained. A natural question is whether a one-off workshop is sufficient to achieve this. Workshop participants themselves raised concerns that the intervention was too short, too isolated, too limited in scale, and lacked follow-up engagement (Appendix Figure K11). Potentially as a consequence of its time-limited nature, the intervention might have been more successful in increasing understanding of problems than agreement over solutions: at endline, there is *less* consensus about priorities for action in treated schools than in control schools.[46] This suggests that the intervention might have successfully generated a range of ideas about possible solutions but not built consensus over a specific and unified action plan. Transparency interventions may be more effective if coupled with longer-term engagement, for example, through a series of workshops in which participants evaluate not only baseline conditions but progress relative to previously identified action plans. Consensus-building over action plans may be as important as consensus-building over problem diagnosis. Additionally, transparency interventions might be more effective if coupled with accountability measures that incentivize adherence to action plans. Consistent with this hypothesis, common features of many successful interventions are a specific focus on creating concrete action plans and a longer time frame that includes follow-up meetings to monitor progress (Table 3a). In contrast, unsuccessful interventions often either fail to bring service providers and users together at all or are characterized by more time-limited interventions (Table 3b).

Challenges of scale For transparency interventions to successfully ignite "action cycles" (Kosack and Fung, 2014), changes made by actors must affect provision. This also implies that interventions must influence actors who have the capacity to affect provision. It is not clear whether this is the case for the AWIS intervention. Semi-structured interview respondents report dependence on external organizations for financial support for major ("hard") infrastructure improvements. But local administrators, who have limited and fragmented responsibility for school WASH services, did not attend the AWIS workshops (see Table 2). In contrast, the actions that students, teachers, and headteachers report taking were primarily "soft" (Appendix Figure K12). Raffler et al. (2020) show suggestive evidence that a similar intervention was more successful when local officials participated in community-service provider workshops, although local officials only participated in about 30% of workshops. Additionally, the only explicit mentions of corruption or trust problems in the semi-structured interviews are with respect to contractors doing shoddy or incomplete work. This class of governance problems is not within the locus of control of school management or leadership.

In contexts like these, transparency interventions may need to think explicitly about how to influence other actors, perhaps using two-stage workshops at different administrative scales. Local administrators might be more likely to attend a district-wide workshop than multiple school-level workshops. Alternatively, transparency interventions might be coupled

[46] Details in Appendix J.

with financial support that frees schools from dependence on external decision-makers. In at least one successful precedent, facilitating collaboration between schools and community leadership increased the impact of a grant (Pradhan et al., 2014).

Misleading expectations In this study, control schools appeared to more successfully improve WASH service provision for female students than treated schools, with the caveat that our experiment was not designed to measure negative effects. However, intervention designers and researchers may want to consider the possibility that stand-alone transparency or accountability interventions might create misleading expectations about future financial support, potentially discouraging participants from trying to solve their own problems in anticipation of this support. In a semi-structured interview, a teacher who participated in the AWIS workshop mentions informing DORP (the local implementing NGO) that the school needed a new well during the AWIS workshop, seemingly with the expectation that DORP would help provide one. While the negative effects we observe appear unique in the published literature, many studies report null effects, and Arkedis et al. (2021) note that expectations about "how these projects work" may shape the choices people make about whether or not to dedicate time and effort to project activities.

The evidence is inconsistent with another potential explanation, that the intervention failed because it antagonized participants or was simply altogether ineffective. The intervention did improve knowledge and institutions. Ex post, survey respondents expressed satisfaction with the intervention, albeit not universally: a substantial share of students and teachers felt that some participants were more influential than others, and about a third of teachers did not feel free to express their opinions freely (Appendix Figure K13). Participants also praised the educational nature of the intervention and the systematic, constructive, and participatory discussion (Appendix Figure K14).

10 Conclusion

This study evaluates the impact of a transparency intervention on WASH service provision in schools in rural Bangladesh. While interventions designed to increase transparency are currently "in vogue", evidence from similar interventions in other sectors finds only mixed effects (Fox, 2015; Joshi, 2010; Kosack and Fung, 2014). The transparency intervention we study is a deliberative multi-stakeholder workshop initiated with a community scoring exercise, closely resembling one of a set of "exemplary transparency interventions" identified by Kosack and Fung (2014). Eight to twelve months after implementation, the intervention improves knowledge of WASH and MHM standards and best practices, and WASH management institutions, but does not improve WASH service provision. Although our study was not designed to detect negative effects, we find suggestive evidence that control schools

improve WASH service provision for female students more successfully than treated schools. We find no detectable effects on school attendance or exam performance.

Drawing on extensive descriptive data, we identify four potential explanations for the absence of a measurable positive effect on WASH service provision despite the positive effects on knowledge and institutions: the relatively short time frame of the study; the time-limited nature of the intervention; the challenges of implementing transparency interventions at a decentralized level when actors are dependent on external sources of support; and the potential for such interventions to create misleading expectations of future support, discouraging stakeholders from independently seeking solutions.

Designers of transparency interventions may wish to consider how to address the potential limitations we outline above. In particular, our results suggest that interventions may need longer time horizons to be effective, for example incorporating repeated workshops to build consensus about concrete action plans and monitor progress. Future studies should also examine how the effects of transparency interventions vary over time.

Bibliography

Ahrens, A., C. Hansen, and M. Schaffer (2020): "lassopack: Model selection and prediction with regularized regression in Stata," *The Stata Journal*, 20.

Alam, M.-U., S. P. Luby, A. K. Halder, K. Islam, A. Opel, A. K. Shoab, P. K. Ghosh, M. Rahman, T. Mahon, and L. Unicomb (2017): "Menstrual hygiene management among Bangladeshi adolescent schoolgirls and risk factors affecting school absence: results from a cross-sectional survey," *BMJ Open*, 7.

Anderson, M. L. (2008): "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American Statistical Association*, 103, 1481–1495.

Arkedis, J., J. Creighton, A. Dixit, A. Fung, S. Kosack, D. Levy, and C. Tolmie (2021): "Can transparency and accountability programs improve health? Experimental evidence from Indonesia and Tanzania," *World Development*, 142.

Balk, D. (1997): "Defying Gender Norms in Rural Bangladesh: A Social Demographic Analysis," *Population Studies*, 51.

Banerjee, A., 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*, 2.

Bangladesh Bureau of Statistics (2020): "National Hygiene Survey 2018," Financial and technical assistance of WaterAid Bangladesh and UNICEF Bangladesh.

Barr, A., F. Mugisha, P. Serneels, and A. Zeitlin (2012): "Information and collective action in community-based monitoring of schools: Field and lab experimental evidence from Uganda," Working paper.

Belloni, A., V. Chernozhukov, and C. Hansen (2014a): "High-Dimensional Methods and Inference on Structural and Treatment Effects," *Journal of Economic Perspectives*, 28.

——— (2014b): "Inference on Treatment Effects after Selection among High-Dimensional Controls," *Review of Economic Studies*, 81.

BenYishay, A., L. Mueller, K. Nolan, and P. Roessler (2022): "Testing multi-stakeholder dialogue for better local governance in Niger: An experiment. Can we talk our way out of development problems?" *Development Policy Review*, 40.

Björkman, M. and J. Svensson (2009): "Power to the People: Evidence from a Randomized Field Experiment on Community-Based Monitoring in Uganda," *Quarterly Journal of Economics*, 124.

Christensen, D., O. Dube, J. Haushofer, B. Siddiqi, and M. Voors (2021): "Building resilient health systems: Experimental evidence from Sierra Leone and the 2014 Ebola outbreak." *Quarterly Journal of Economics*, 136.

Clasen, T., D. Fabini, S. Boisson, J. Taneja, J. Song, E. Aichinger, A. Bui, S. Dadashi, W.-P. Schmidt, Z. Burt, and K. L. Nelson (2012): "Making Sanitation Count: Developing and Testing a Device for Assessing Latrine Use in Low-Income Settings," *Environmental Science & Technology*, 46.

Cohen, J. (1988): *Statistical power analysis for the behavioral sciences*, Hillsdale, NJ: Erlbaum, 2nd ed.

Djimeu, E. W. and D.-G. Houndolo (2016): *Power calculation for causal inference in social science: Sample size and minimum detectable effect determination. 3ie Working Paper 26.*, New Delhi: International Initiative for Impact Evaluation (3ie).

Duflo, E. (2018): "Machinistas meet randomistas: useful ML tools for empirical researchers," NBER Summer Institute.

Fabbri, C., V. Dutt, V. Shukla, K. Singh, N. Shah, and T. Powell-Jackson (2019): "The effect of report cards on the coverage of maternal and neonatal health care: a factorial, cluster-randomised controlled trial in Uttar Pradesh, India," *Lancet Global Health*, 7.

Flug, D., R. Largo, and A. Prader (1984): "Menstrual patterns in adolescent Swiss girls: a longitudinal study," *Annals of Human Biology*, 11.

Fox, J. A. (2015): "Social Accountability: What Does the Evidence Really Say?" *World Development*, 72.

Gelman, A. (2019): "Don't Calculate Post-hoc Power Using Observed Estimate of Effect Size," *Annals of Surgery*, 269.

Gibbons, C. E., J. C. S. Serrato, and M. B. Urbancic (2019): "Broken or Fixed Effects?" *Journal of Econometric Methods*, 8.

Government of Bangladesh (2011): "National Strategy for Water and Sanitation Hard to Reach Areas of Bangladesh," .

ICRRDB, Water Aid, and Policy Support Unit (2014): "Bangladesh National Hygiene Baseline Survey," Preliminary report.

Imbens, G. W. and D. B. Rubin (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press.

J-PAL (2018): *Reducing costs to increase school participation*, J-PAL Policy Insights, Abdul Latif Jameel Poverty Action Lab.

Joshi, A. (2010): "Annex 1: Service Delivery," in *The Impact and Effectiveness of Transparency and Accountability Initiatives: A review of the evidence to date. Synthesis Report*, ed. by R. McGee and J. Gaventa, Transparency Accountability Initiative.

Kosack, S., G. Bridgman, J. Creighton, C. Tolmie, P. Whitt, and A. Fung (2019): "Encouraging Participation," Working paper.

Kosack, S. and A. Fung (2014): "Does Transparency Improve Governance?" *Annual Review of Political Science*, 17.

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.

Lin, W. (2013): "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique," *Annals of Applied Statistics*, 7.

McMichael, C. (2019): "Water, Sanitation and Hygiene (WASH) in Schools in Low-Income Countries: A Review of Evidence of Impact," *International Journal of Environmental Research and Public Health*, 16.

Miguel, E. and M. Kremer (2004): "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," *Econometrica*, 72, 159–217.

Mohanan, M., V. S. Rajan, K. M. Swanson, and H. Thirumurthy (2020): "Information and Facilitation Interventions for Accountability in Health and Nutrition: Evidence from a Randomized Evaluation in India," Economic Research Initiatives at Duke (ERID) Working Paper 295.

Nyqvist, M. B., D. de Walque, and J. Svensson (2017): "Experimental Evidence on the Long-Run Impact of Community-Based Monitoring," *American Economic Journal: Applied Economics*, 9.

OECD (2015): *OECD Principles on Water Governance*, OECD.

Olken, B. A. (2007): "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy*, 115.

Pandey, P., S. Goyal, and V. Sundararaman (2009): "Community participation in public schools: impact of information campaigns in three Indian states," *Education Economics*, 17.

Pandey, P., A. R. Sehgal, M. Riboud, and D. L. M. Goyal (2007): "Informing Resource-Poor Populations and the Delivery of Entitled Health and Social Services in Rural India: A Cluster Randomized Controlled Trial," *Journal of the American Medical Association*, 298.

Pradhan, M., D. Suryadarma, A. Beatty, M. Wong, A. Gaduh, A. Alisjahbana, and R. P. Artha (2014): "Improving Educational Quality through Enhancing Community Participation: Results from a Randomized Field Experiment in Indonesia," *American Economic Journal: Applied Economics*, 6.

Raffler, P., D. N. Posner, and D. Parkerson (2020): "Can Citizen Pressure be Induced to Improve Public Service Provision?" Working paper.

Roby, D. E. (2004): "Research on school attendance and student achievement: A study of Ohio schools." *Educational Research Quarterly*, 28.

Samii, C. (2018): "make_index," Github repository.

Tanjeela, M. (2021): "Female leadership in local governments: reconstruction of traditional gender identity of women in rural Bangladesh," *Pakistan Journal of Women's Studies*, 28.

UNDP (2011): *Fighting Corruption in the water Sector: Methods, tools and good Practices*, United Nations Development Programme.

UNESCO (2014): "Puberty Education and Menstrual Hygiene Management," *Good Policy and Practice in Health Education - Booklet 9*, 1–59.

UNICEF (2012): *Water, Sanitation and Hygiene (WASH) in Schools: A companion to the Child Friendly Schools Manual*, UNICEF Division of Communication, New York.

Water Aid (2016): "School WASH research: Bangladesh country report," Report prepared for Water Aid by Jacques-Edouard Tiberghien, Partnerships in Practice.

Westfall, P. H. and S. S. Young (1993): *Resampling-based multiple testing: Examples and methods for p-value adjustment*, Wiley.

WIN and IRC (2011): *Annotated Water Integrity Scans: A manual to help assess integrity levels in specific sub-sectors of the water sector*, Water Integrity Network and IRC International Water and Sanitation Centre.

Young, A. (2019): "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results," *Quarterly Journal of Economics*, 134.

We evaluate a deliberative stakeholder workshop initiated with a community scorecard.

Our experiment evaluates effects on WASH provision in schools in rural Bangladesh.

The intervention leads to moderate improvements in knowledge and institutions.

The intervention does not improve service provision or change use patterns.

Descriptive data suggest three potential explanations.

Umrbek Allakulov: conceptualization, funding acquisition, writing – review and editing. Serena Cocciolo: conceptualization, methodology, software, formal analysis, data curation, writing – review and editing. Binayak Das: conceptualization, funding acquisition, writing – review and editing. Md. Ahasan Habib: conceptualization, methodology, investigation, data curation, supervision, project administration, writing – review and editing. Lovisa Rambjer: methodology, software, formal analysis, data curation, visualization, writing – review and editing. Anna Tompsett: conceptualization, methodology, software, formal analysis, resources, writing – original draft, writing – review and editing, visualization, supervision.