MCC – Lesotho
Rural Water Supply and Sanitation Activity

Impact Evaluation Report

September 20, 2017

NORC
at the UNIVERSITY of CHICAGO

TABLE OF CONTENTS

TABLE OF TABLES

ABBREVIATIONS

AMP	Activity Monitoring Plan
BOS	Lesotho Bureau of Statistics
CLO	Community Liaison Officer
CTV	Continuous treatment variable
DD	Difference in differences
DHS	Department of Health Services
DRWS	Department of Rural Water and Sanitation
DQA	Data Quality Assessment
EA	Census enumeration area
GOL	Government of Lesotho
HALEH	Household Awareness Latrine and Environmental Hygiene
HH	Household
IEMS	Impact Evaluation Multipurpose Surveys
IFB	Information for Bid
ITT	Indicator Tracking Table
LMDA	Lesotho Millennium Development Agency
MCA	Millennium Challenge Account
M&E	Monitoring and Evaluation
MDES	Minimum detectable effect size
NORC	NORC at the University of Chicago
O&M	Operations and maintenance
PDNA	Engineering company responsible for Urban Water oversight and reporting
PHAST	Participatory Health and Sanitation Training
PMCS	Project Management and Construction Supervision
PSU	Primary statistical unit
RCT	Randomized control trial
RDD	Regression discontinuity analysis
RSA	Researcher Supplemental Application
RWSSA	Rural Water Supply and Sanitation Activity
TBD	To be determined
TOR	Terms of reference
ToT	Treatment of the treated

VIP	Ventilated Improved Pit
VWHC	Village Water and Health Committee
WASCO	Water and Sanitation Company
WHO	World Health Organization
WM	Water Minder
WSP	Water and Sanitation Project
WBD	Water-Borne Disease

Page intentionally left blank

EXECUTIVE SUMMARY

The Millennium Challenge Corporation (MCC), through its Compact with the government, awarded $164-million over five years for investment in improved water supplies and sanitation facilities for rural and urban domestic, commercial, and industrial users. As part of its commitment to transparently and thoroughly monitor and evaluate its activities, the MCC contracted NORC in 2007 to conduct an impact evaluation of its water sector activities. This report presents the Impact Evaluation of the Rural Water Supply and Sanitation Activity (RWSSA).

RWSSA originally included 250 rural water supply points and 10,000 VIP latrines and had a budget of $30.2 million (18 percent of the $164-million Water Project in the Compact). In order to increase the coverage of VIP latrines in participating villages, MCC subsequently increased the budget to $40.1 million and the Government of Lesotho (GOL) contributed $17.1 million to RWSSA.[1] In addition, the target for VIP latrines coverage was increased from 10,000 to 27,245 in the Lesotho M&E Plan. When the Lesotho Compact ended in September 2013, 175 water systems (70% of the target) and 29,352 VIP latrines (108% of the target) had been installed.[2]

Implementation continued post-Compact with approximately $5.3 million of additional funding from the GOL; ultimately, 250 water systems (100% of the target), and 31,768 VIP latrines (117% of the revised target), were completed.[3] The total cost of RWSSA, including MCC and GOL funding during the Compact and after, was approximately $60 million.[4]

Households impacted by the program are located in villages that were identified by the Department of Rural Water Supply (DRWS) as lacking access to safe drinking water and adequate sanitation. To identify the effects of the program on the outcomes of interest, eligible villages were randomly assigned to treatment and control groups. The list of outcomes analyzed includes toilet use, type of water source used, time spent collecting water, diarrhea incidence and income, among others.

To evaluate the impact of the program we use data from the baseline and follow-up Impact Evaluation Multipurpose Surveys (IEMS). The IEMS is a longitudinal analytic survey specifically designed to collect data for the impact evaluations of the MCA-Lesotho Compact health and water (rural and urban) activities.

During program implementation construction delays in some treatment villages prevented that construction works ended before follow-up data collection

[1] This was a net amount; the GOL actually contributed $32.3 million but was reimbursed $15.1 million by MCC (Source: Office of Inspector General, 2014; and Lesotho Millennium Development Agency, 2014).

[2] Source: Lesotho Compact Project Results, MCC website. Retrieved from https://www.mcc.gov/where-we-work/program/lesotho-compact

[3] Source: Own calculations using data from Cowater International Inc. (2016), Lesotho Millennium Development Agency (2014) and Lesotho Millennium Development Agency (2015).

[4] In addition, the GOL expanded the scope further to supplement some of the systems constructed under the Compact and also build 19 water systems and 4,554 VIP latrines in neighboring communities using separate funds (Source: Cowater International Inc., 2016).

were a subset of the villages that were assigned to the treatment group originally. To tackle this problem we used Instrumental Variables (IV) methods in order to evaluate the effect of the program. This approach exploits the fact that treatment assignment was randomized, but it also addresses the fact that treatment was not provided as planned in all treatment villages.

The impact evaluation shows that the program has had significant effects on key wellbeing indicators. We found that households in the treatment group are more likely than in the control group to use as their main water source an improved water source, such as a public standpipe or a protected spring, as opposed to an unimproved source, such as an unprotected spring or surface water. They are also more likely to use a toilet and spend less time collecting water. However, we did not find any impacts significant at standard levels of confidence for any of the diarrhea incidence indicators we analyzed, although most of the estimated effects have the (negative) expected sign.

We also did not find any effects for any labor outcomes, or income. An important exception to this is that we found that the program has a positive and significant effect on female labor participation. We discuss the mechanisms that can explain why the effects on labor outcomes are not more apparent. In particular, it is possible that time availability does not translate into better labor outcomes because the latter are not restricted by time availability but by other conditions, like the labor market itself.

In terms of policy implications, the results described in this report imply that this type of program can have major impacts on households wellbeing via reductions on time spent collecting water, but limited effects on higher level outcomes, like diarrhea incidence. Furthermore, even if household members spend less time collecting water as a result of the program, it is not clear that this will translate into a 1:1 increase in the number of hours they participate in the labor market, as labor outcomes may depend on more factors than just greater available time.

1. INTRODUCTION

Improving access to safe drinking water and basic sanitation can bring health, social and economic benefits. For example, it is estimated that in Africa, people spend 40 billion hours every year collecting water.[5] In order to help realize these benefits in Lesotho, the Millennium Challenge Corporation (MCC), through its Compact with the government, awarded $164-million over five years for investment in improved water supplies and sanitation facilities for rural and urban domestic, commercial, and industrial users.

As part of its commitment to transparently and thoroughly monitor and evaluate its activities, the MCC contracted NORC in 2007 to conduct an impact evaluation of its water sector activities. This report presents the Impact Evaluation of the Rural Water Supply and Sanitation Activity (RWSSA).

In aiming to reach the Millennium Development Goals (MDG) by 2015, Lesotho has faced a particular challenge in improving rural water delivery, which has remained relatively static at 75-77 percent in the 25 years between 1990 and 2015. Over the same period, there has been an even greater challenge in improved sanitation, even though rural coverage has grown from 20 to 28 percent.[6] Given that 95 percent of households in urban areas have improved water sources, a major thrust for improved services was clearly needed in the rural sector.

Despite numerous investments both by donor agencies and the government to increase access to water sources and sanitation facilities, a recent assessment of progress towards MDG goals by the Joint Monitoring Project (JMP) found that, in Lesotho, there has been "limited or no progress" in sanitation and "moderate progress" in water coverage. In both sectors Lesotho was regarded as not having met the MDG targets. [7]

The marked discrepancy in the relatively high rates of water coverage and the low levels of sanitation coverage also points to an additional challenge that is particularly prevalent in rural areas. A strategy to reduce water-borne disease (WBD) through water and sanitation must ensure that hygiene promotion, sanitation facilities and water systems are combined to achieve the desired impact.[8] Often, program attention tends to focus on the delivery of water systems without simultaneous attention being paid to sanitation facilities and hygiene promotion; by contrast, the RWSSA included all three components, hopefully setting the stage for long-term impacts in reducing disease and improving the productive lives of Lesotho's citizens.

[5] www.charitywater.org/whywater/

[6] Estimates on the Use of Water Sources and Sanitation Facilities, Lesotho, updated June 2015. WHO/UNICEF Joint Monitoring Program for Water Supply and Sanitation.

[7] UNICEF/WHO. Progress on Sanitation and Drinking Water – 2015 update and MDG assessment. 2015. Annex 3. http://www.wssinfo.org/fileadmin/user_upload/resources/JMP-Update-report-2015_English.pdf

[8] Cairncross, Sandy, et al. Water, sanitation and hygiene for the prevention of diarrhea. International Journal of Epidemiology 2010;39: 193–205.

2. OVERVIEW OF THE COMPACT AND THE INTERVENTIONS EVALUATED

2.1 Overview of the Project Implementation

2.1.1 Program Description

RWSSA originally included 250 rural water supply points and 10,000 VIP latrines and had a budget of $30.2 million (18 percent of the $164-million Water Project in the Compact). In order to increase the coverage of VIP latrines in participating villages, MCC subsequently increased the budget to $40.1 million and the Government of Lesotho (GOL) contributed $17.1 million to RWSSA.[9] In addition, the target for VIP latrines coverage was increased from 10,000 to 27,245 in the Lesotho M&E Plan. When the Lesotho Compact ended in September 2013, 175 water systems (70% of the target) and 29,352 VIP latrines (108% of the target) had been installed.[10]

Implementation continued post-Compact with approximately $5.3 million of additional funding from the GOL; ultimately, 250 water systems (100% of the target), and 31,768 VIP latrines (117% of the revised target), were completed.[11] The total cost of RWSSA, including MCC and GOL funding during the Compact and after, was approximately $60 million.[12]

Water system modalities included boreholes with hand pumps, solar powered pumping systems, gravity-fed spring catchment systems, and electric pumping systems. Each system encompassed between 2 and 5 villages. In treated villages standpipes were placed according to the village's demand and how far apart houses were from each other. According to the audit by the Project Management and Construction Supervision, for the most part households in treated villages had a standpipe within 150m of distance, which is the DRWS standard for service.[13]

In addition to MCC-funded construction of new water systems and VIP latrines, DRWS also provided Participatory Hygiene Awareness and Sanitation Training (PHAST) and Aftercare training to participating villages.[14] PHAST, which occurred before the construction of water systems commenced, consisted of two components:

[9] This was a net amount; the GOL actually contributed $32.3 million but was reimbursed $15.1 million by MCC (Source: Office of Inspector General, 2014; and Lesotho Millennium Development Agency, 2014).

[10] Source: Lesotho Compact Project Results, MCC website. Retrieved from https://www.mcc.gov/where-we-work/program/lesotho-compact

[11] Source: Own calculations using data from Cowater International Inc. (2016), Lesotho Millennium Development Agency (2014) and Lesotho Millennium Development Agency (2015).

[12] In addition, the GOL expanded the scope further to supplement some of the systems constructed under the Compact and also build 19 water systems and 4,554 VIP latrines in neighboring communities using separate funds (Source: Cowater International Inc., 2016).

[13] Source: Cowater International Inc. (2016).

[14] PHAST and Aftercare trainings were jointly funded by DRWS. MCA provided funds to DRWS to provide snacks to community members and lunches for the VHWC during PHAST, as well as per diems for the Community Liaison Officer (CLO) to cover the cost of paying a village household for lodging. DRWS paid the CLO their salary and used government resources (car and petrol) to get to and from the village.

- *Community-wide hygiene awareness and sanitation training*: Delivered to the entire community by a Community Liaison Officer (CLO), this training consists of a participatory approach in which the CLO conducts a transact walk through the village with the entire community. During this walk, the CLO raises awareness about hygiene and sanitation by pointing out examples unhygienic/unsanitary practices and informing them about solutions they must implement to change those practices and improve hygiene within the community.
- *Training to Village Water and Health Committees (VWHCs):* After the community-wide PHAST was completed, the community democratically elected the VWHC. The role of the VWHC, of which the Water Minder is a member, is to serve as a source of information to the community on benefits of access to clean water, water-related disease control, disposal of dirty water/waste water management, and types of latrines and their requirements, among other hygiene and sanitation topics.[15] CLOs, using a series of pictures (about 80 pictures with Sesotho script), trained the VWHC on good hygiene practices. The VWHC members were also tasked with helping community members build their hand-washing models (tippy-taps, for example) and soak away pits, and informing them of preparations required to receive a VIP latrine. Towards this end, the CLOs train VWHC members on positive hygiene and sanitation practices and teach them how to build hand-washing models/tippy taps. The VWHC treasurer also received training on keeping an account book.

As mentioned before, PHAST generally took place in the pre-construction phase, and served as one indicator of a village's "readiness" for construction of a water system. In most of the villages, VWHC training also took place before construction began. PHAST training was the responsibility of DRWS.

Aftercare Training: In addition to PHAST training, DRWS was also responsible for providing Aftercare Training to VWHCs. In keeping with the World Bank strategy on water supply [16], the DRWS Aftercare Strategy aims to put in place institutional and financial mechanisms to sustain the construction of water supplies for their 10-15-year design life. Aftercare training occurred after the construction of the village Water & Sanitation System. As described in the DRWS Community Management Handbook, this training was intended to build the VWHC capacity to perform all operation and maintenance activities on the village water and hygiene system.

Separate from the DRWS Aftercare Training, all village Water-Minders were also supposed to receive on-site training from the building contractor to learn to operate their village water system. This training should have occurred during the construction of the water system. Water Minders were expected to participate in the construction process, so that they are well informed about the make-up of the system. After the completion of the system, the Water Minder was supposed to receive a copy of the Operation & Maintenance Manual for the water system written by the contractor, as well as a toolkit for maintenance functions.

[15] The Village Water Minder is a member of the VWHC. His/her primary responsibility within the committee is to identify and report maintenance problems that require attention to the VWHC. The Water Minder also presents to the committee a cost estimate for fixing the problem at hand, so that the VWHC treasurer can provide her/him with the required funding to buy parts and repair the system.

[16] World Bank's Water Global Practice – Strategy
http://www.worldbank.org/en/topic/water/overview#2

2.1.3 Project Implementation

The Lesotho RWSSA was implemented by DRWS in all 10 districts of the country. DRWS selected 250 water projects to provide services in villages that lacked access to safe drinking water and adequate sanitation.[17] This selection of projects was made from lists of villages in which ready-to-implement projects would reside that district representatives provided.[18] In August 2008, NORC facilitated the random assignment of 100 of the 250 water projects into groups targeted for treatment and control. The random assignment of the 100 water systems to treatment and control status was conducted in a public event to assure transparency.

From each of these selected water projects, the village in which the system would reside was placed into a corresponding treatment or control group for the purposes of the impact evaluation.[19] Thus, while a given water project would provide similar household-level benefits to more than one village, for data collection purposes we selected just the village in which the water system resides, not water systems, as the *de facto* PSU.[20] From these 100 associated villages within which the water systems were to reside, 50 villages (5 in each district) were randomly selected for the first wave of project implementation (Phase A); this group constituted the treatment group. The remaining 50 villages from the 10 districts were assigned to the control group (Phase C). The remaining 150 villages (or, technically, their water systems) did not constitute part of the evaluation sample. Annex A provides a map of the treatment and control villages used in the present study.

The construction of water systems in the 50 Phase-A treatment villages commenced between December 2010 and March 2011. Although they were scheduled to be completed by September 2011, there were several delays in the construction schedule. Most importantly, in April 2012, 13 to 16 months after construction began, the Millennium Challenge Account (MCA) terminated three contracts of the construction companies responsible for building the water systems in 11 Phase-A treatment villages. One year later, in April 2013, new contractors took over in these 11 treatment sites, and continued the interrupted construction process. For ease of presentation and differentiation, we will refer to the group of 39 villages that continued construction with no contractual disruptions as Phase-A^{rev} villages, and to the remaining 11 villages as Phase A_1.[21]

Table 1 below presents construction start and end dates, and PHAST community training dates for villages in Phases A and C. The *official construction completion date* represents the point at which all construction activities of water and sanitation structures (water systems and VIP

[17] DRWS also applied additional criteria, including the quality of village governance and enthusiasm for the infrastructure.

[18] While a water project would reside in a given village, in many cases a system provided water and sanitation services to more than one village.

[19] NORC at the University of Chicago. Impact Evaluation Design & Implementation Services – Lesotho. Evaluation Mini-Report. January 2009.

[20] Thus, in principle, the evaluation is relevant only for villages in which the water system resides. However, according to Sello Sefali (LMDA), given that treatment for the household simply comprised availability of a standpipe with clean water, the impacts of the intervention in the other villages *that are part of the same water system* should be identical.

[21] In DRWS and MCA documentation, the original treatment group is referred to as "Phase A." However, following the splintering off of 11 Phase-A_1 villages, the original treatment group *minus* the Phase-A_1 villages was also referred to as "Phase A". In this Report, to avoid confusion, we refer to the original 50 Phase-A villages as Phase A and the reduced set of 39 Phase-A villages (the set minus the delayed Phase-A_1 villages) as Phase A^{rev}.

latrines) were completed, inspected and certified; on this date, after all defects and problems had been rectified by the contractor, the engineer issued the village a Certificate of Completion (CoC). Table 1 also presents the *approximate physical construction completion date prior to inspection*.

Construction of Phase C control villages commenced between January and April 2013.

Table 1. Construction and Training Dates for Phase Arev, A1, and C Villages

Phase	Type	# villages	Dates (number of villages)					
			Start of construction*	Approximate completion of construction, prior to inspection*	Official construction end date (CoC issued)*	PHAST community training†	VWHC training†	Aftercare training§
A^{rev}	Treatment (T1)	39	Dec 2010 – Feb 2011 (39)	Jun 2011 – Mar 2012 (34) Oct 2012 – Dec 2012 (3) Sep 2013 (2)	Oct 2011 (2) Dec 2011 – Feb 2012 (19) Aug 2012 (13) Sep 2013 (5)	Feb 2008 (1) Apr 2008 (1) Sep– Dec 2008 (26) Mar 2009 – Nov 2009 (10) Nov 2010 (1) May 2013 (1)	Jan 2010 – Nov 2010 (37) May 2013 (2)	Aug 2012 (2) Apr 2013 – Jun 2013 (16) Nov 2013 - Oct 2013 (2)
A_1	Treatment (T2)	11	Jan 2011 – Mar 2011 (11)	Jul 2013 – Nov 2013 (9) Feb 2014 (2)	Sep 2013 (4) Jan 2014 (4) Mar 2014 (1) Aug 2014 (2)	Apr 2008 – Dec 2008 (7) Apr 2009 (2) Nov 2010 (2)	Jul 2010 (1) Oct 2010 – Nov 2010 (10)	-
C	Control	48	Jan 2013 – Apr 2013 (48)	Mar 2013 (1) May 2013 – Jun 2013 (7) Aug 2013 – Dec 2013 (39) *Unknown (1)*	Sep 2013 – Mar 2014 (46) *Unknown (2)*	Mar 2009 (1) Jan 2010 – Oct 2010 (17) Mar 2011 (1) Nov 2011 (3) Feb 2012 – Nov 2012 (25) Mar 2013 (1)	Jan 2012 (1) Mar 2012 (7) Oct 2012 – Jul 2013 (39) Nov 2013 (1)	Feb 2013 (2) Nov 2013 (2) Jan 2014 (7) Feb 2014 (2)

Sources: * MCC's supervisory engineer, data from Cowater monthly reports
† Cowater / NORC
§ Cowater / NORC. Data reflects reality in 2013-2014. DRWS may have provided aftercare to more villages since then.

The construction of water systems in all but five of the 39 Phase A^{rev} treatment villages was physically completed between June 2011 and March 2012, which indicates a minimum 10-month treatment period before construction began in Phase-C villages. Three others were completed in the last quarter of 2012, and two were delayed until September 2013. Completion of construction in all Phase-A_1 villages was severely delayed with nine completing physical construction in Jul-Nov 2013, and two in February 2014. Hence, of the group of 50 treatment villages, construction in all 11 Phase-A_1 villages and one Phase A^{rev} was completed only after construction in Phase-C villages had already begun.

Early on in the Compact, the Government of Lesotho supplemented MCA investments to enable every household in a treatment village to receive a VIP latrine. The 100-percent sanitation coverage plans were based on listings of households within the village conducted by DRWS. For various reasons – quality issues during the listing, which resulted in some households being missed, and a lag between listing and start of construction, during which new households were built – some households in Phase-A villages did not receive VIP latrines. As we discuss more thoroughly in the following sections, to measure which households do have a VIP latrine at follow-up, we rely on data form the Impact Evaluation Multipurpose Surveys (IEMS).

A second consequence of the 100-percent VIP latrine coverage plan is that there is no way for the evaluation to disentangle the separate contributions of the water system and VIP access. Also, since treatment and control villages both received PHAST training prior to the baseline survey, this means that the design cannot isolate the contribution of PHAST to any outcomes.

2.2 Program Logic for the Rural Water Project

MCC's new program logic diagram (see Annex B) for the Lesotho RWSSA, shared with NORC in March 2013, presents activities and outputs that are linked to short-, intermediate- and long-term types of outcomes.[22] These effects are:

Short-term outcomes

1. Increased hygiene awareness among communities
2. Increased access to improved sanitation
3. Increased access to improved water sources
4. Increased awareness/knowledge of Water Committees, Water Minders, and communities in maintaining systems

Intermediate outcomes

5. Improved hygiene behavior[23]
6. Decreased water-related illness
7. Reduced expenditure on medical care

[22] While NORC reviewed the ERR in addition to the program logic and the literature when establishing the research hypotheses with MCC, NORC was informed by Jennifer Sturdy (DPE/EE-ME) in an email of April 23, 2014 that she had spoken with MCC's economist for Lesotho, Sarah Olmstead, and "she confirms that for this Compact and Rural Water specifically, there is no need to link to the ERR and Beneficiary Analysis."

[23] An MCC reviewer stated that "Improved hygiene behavior is too broad to be an outcome – what is intended here? Hand Washing? Chlorine treatment for water?" This outcome, however, is taken directly from the MCC Program Logic, which includes "Improved hygiene behavior" as an Intermediate Outcome.

8. Time saved in water collection
9. Maintenance of systems by Water Minders

Long-term outcomes

10. Increased productive activity (productivity)
11. Increased income

2.3 Evaluation Hypotheses and Indicators

The evaluation hypotheses for the impact evaluation are linked to the outcomes presented above and in MCC's program logic. Table 2 presents the following information:

- Maps out the evaluation hypotheses related to the rural water-supply investments to key outcome/impact indicators (Columns 1 and 2);
- Indicates the minimum time of exposure necessary to detect changes in the outcome indicators (Column 3);
- Maps outcome indicators to treatment indicators, as shown in the pathways in MCC's Program Logic (Column 4).

Note that not all outcomes considered in the program logic (Section 2.2) are included in Table 2 due to methodological considerations.[24]

[24] Five of the 13 hypotheses in the Revised Evaluation Design Report were supposed to be tested using Continuous Treatment Variable (CTV) approach. After discussing with MCC's Evaluation Management Committee (EMC) it was decided that it was preferable to drop the hypotheses that use this method and focus on the ones that could be evaluated using randomization as the key source of variation for identification of the treatment effect. As a result, no analysis using CTV is presented.

Table 2. Rural Water Project: Reviewed Mapping of Evaluation hypotheses to Impact Indicators, Treatment Indicators and Data Sources

Evaluation period / Hypothesis	Outcome or Impact Indicator		Treatment Indicator[a]
	Description	Months required	
Short-term outcomes			
1. Access to improved water systems increases household use of safe drinking water	Degree to which household collects water from improved sources	3 months	Water system constructed
2. Installation of a VIP increases use of improved sanitation	Frequency of household use of VIP latrine	3 months	VIP latrine constructed
3. Access to improved water source reduces time spent collecting water [b]	Time (minutes) spent collecting water	3 months	Water system constructed
Intermediate Outcomes			
4. Program reduces incidence of WRD.	Incidence of diarrhea over last 2 weeks	6 months	Water system constructed
5. Program reduces expenditure on medical care	Household medical expenditures	6-9 months	Water system constructed
6. Program increases school attendance	# of days that school-age children in household attend school in past two months	6 months	Water system constructed
7. Program provides time-savings to house-holds associated with better health status	Time (days) lost from illness Time (days) lost in caring for the sick	6 months	Water system constructed
Long-term outcomes			
8. Access to improved water system, VIP, PHAST lead to increased household income	Hours worked Income (of household as a whole & indi-vidual members)	9 months	Water system constructed

(a) The treatment variable has changed for some hypotheses with respect to what was presented in the Revised Evaluation Design Report. In particular, in some cases the idea was to explore the effect of the program though different channels, (e.g. time savings explaining school assistance), so the treatment variables that needed to be used where continuous. MCC's EMC requested that NORC abstain from using the CTV methods; hence, we dropped all analysis that required this method, and focused on the effect of the program as a whole rather than this type of specific channels. Also, note that the treatment indicator column refers either to when the water systems were constructed or availability of VIP latrines. Because we only have one instrument (the randomized treatment assignment), we can only estimate the effect of one program component at a time (see section 4.2 for a detailed discussion on this issue).

(b) Although "time saved in water collection" is presented as an intermediate outcome in MCC's Program Logic, we anticipate changes to occur in a 3 month time frame. As such, we include it here as a short-term outcome.

Note: Cowater Monthly Progress Reports is the data source for all treatment indicators except VIP latrine constructed. IEMS is the data sources for all outcome indicators and the VIP latrine treatment indicator.

3. LITERATURE REVIEW

The benefits of WASH (Water, Sanitation, and Hygiene) programs are often cited. Meta-analysis and systematic reviews (such as Fewtrell *et al.*, 2005) found water, sanitation, and hygiene interventions to reduce significantly the risks of illness such as diarrhea illness. In terms of the benefits of improved water quality specifically, there is wide consensus in the research of the positive and significant health benefits, in both meta-analysis and systematic review (Esrey *et. al.*, 1999) and in relevant studies in rural areas (see Annex C). Safe drinking water improves health largely by reducing occurrence of diarrhea, a very common illness in the developing world, and other water-related illness. The largest health gains, especially in terms of mortality, are to children under five (see Annex C).

In regard to the hygiene and sanitation component of the program, there is also evidence in the literature of health benefits of such programs, and there is some evidence that all of the WASH interventions more effective when combined. Improved health, in turn, should lead to a number of benefits, including reduced medical costs, reduced time seeking medical care (which can therefore can lead to more time spent at productive income-generating activity), and improved productivity (which should lead to improved wages or outputs per hour).

The literature also indicates that improved access to water reduces water collection time, releasing time and resources for productive activities, such as work and school. However, the data on the amount of time saved is scarcer. Nonetheless, we highlight some in Annex C. We also highlight the literature on the longer term impacts of the program, such as increased productivity, school attendance, and ultimately, income.

The impact evaluation for the RWSSA will provide experimental estimates of the effect of a water and sanitation intervention on a wide set of indicators including diarrhea incidence, time savings, and income. Perhaps the most interesting contribution is the analysis on time savings and its implications on labor outcomes, as this type of mechanisms are less documented in the literature than, for example, diarrhea incidence.

4. DATA COLLECTION AND PROCESSING

The data sources for this evaluation are the baseline and follow-up Impact Evaluation Multipurpose Surveys (IEMS).[25] The IEMS is a longitudinal analytic survey specifically designed to collect data for the impact evaluations of the MCA-Lesotho Compact health and water (rural and urban) activities.

4.1 Sample Frame and Sample Design

The sampling frame for the IEMS consists of all villages in Lesotho based on publicly available geospatial data and 2006 Census data. Information on administrative location, geo-coordinates, rural-versus urban designation and population was merged with publicly available physiographic and geographic data to be used as covariates in the sampling. From this central dataset, sample frames were designed and PSUs were selected for the water (rural and urban) and health project components. For rural water, villages were the primary sampling units.

The sample selection was sequentially sampled without replacement in the form of a two-stage cluster design for the rural water intervention. They cover the designation and selection of villages (PSUs, clusters) and households (SSUs):

Village sample. As described in greater detail in Section 2.1.3, from the 250 water systems in 10 districts selected by DRWS for the MCA rural water interventions, 100 water systems (10 per district) were deemed "ready" for the intervention in 2008. The village in each of the 100 water systems resided was sampled for IEMS. Fifty of these 100 villages were randomly assigned to treatment (Phase A), while the remaining 50 were assigned to the control group (Phase C). Final implementation lists, however, consisted of 50 treatment villages, but only 47 control villages. The village locations are shown on the map in Annex A.

Household sample. Within each treatment and control village a systematic random sample of 13 households was selected.[26] The interview was conducted with the head of the household or the person in the household most knowledgeable about household water and sanitation issues.

[25] The original evaluation design contemplated a third round of data collection to explore the trajectory of results over time. However, having reviewed the follow-up results, NORC and MCC agreed that an additional round of data collection was unlikely to improve our understanding of the program impacts measured at follow-up.

[26] Prior to conducting the IEMS, BoS had conducted a listing of all of their Enumeration Areas. Each EA consists of several of villages. The listing of households within an EA starts from the outer edge of this cluster of villages at a recognizable structure such as a church, store, or health facility identified by BoS' GIS team. The northernmost household to that structure is listed as Household 1 within the EA area. Then all other households are numbered in order in a clockwise direction starting from the outer circumference and moving inward in a circular fashion throughout the entire EA. The final stopping point of the listing is the last household at the very center of the EA. IEMS required the sampling of villages. To sample from each villages, BoS organized each village's household list in numerical order in excel. From each of the village household lists, BoS utilized Excel's RAND function to select a random starting point at which to begin systematically sampling from the village lists. An appropriate sample interval was selected according to the number of households within a village and systematic sampling was carried out to obtain the required sample size for each of the IEMS villages.

4.2 Data Collection

The baseline IEMS was conducted in December 2010, prior to the start of the construction of water and sanitation systems that occurred between December 2010 and March 2011 in treatment villages. As evident from Table 1, however, PHAST training in the vast majority of Phase-A and some Phase-C villages preceded the baseline data collection. Therefore, the December 2010 data collection only serves as a true pre-intervention baseline for the construction of water and sanitation systems. The baseline data collection covered treatment and control (Phase A and Phase C) villages for the rural water intervention. It also covered villages and enumeration areas for the urban water and health sector activities. In November-December 2012, BoS conducted a follow-up data collection. The objective was to collect panel data from the sample of households from the baseline.

As described in more detail in the Revised Evaluation Design Report (Revised: February 23, 2015) data collection by BoS suffered mishaps in both baseline and follow-up data collections. Concerning the baseline, there were delays by BoS in revisiting the field to rectify improper execution of disposition coding, which may have implications for bias in variables of interest. At the follow-up, for unexplained reasons, in 75 villages BoS ignored the fact that they were collecting panel data and interviewed new households instead of returning to the same households as for baseline. As a result, BoS had to return to the field in April 2013 to interview the missing baseline households. This fragmentation of the follow-up data collection poses threats to the evaluation design and may threaten its internal validity.

4.3 Data Processing

Both the baseline and follow-up datasets underwent extensive data consistency checks and cleaning procedures prior to merging. These largely consisted of checking if logical skips in the questionnaire were correctly followed, and making adjustments to the data accordingly. Out-of-range responses were corrected or changed to missing values.

Also before merging, the generated indicators used for analysis were calculated separately in each dataset. In most cases, the survey questions underlying each indicator were the same in both rounds, but in some other cases, differences in the instruments required different formulae.[27]

Originally, cases were to be matched using a unique household ID, comprising the BoS enumeration area (EA) code, and a two-digit suffix representing the order of the household within the EA. However, for most of the IEMS sample, the primary sampling unit was the village, rather than EA, making this an inadequate method of matching panel cases; duplicates are rife (that is, two households had the same ID number), and the precise boundaries of EAs are not always well known in rural areas.

Therefore, a unique case identifier had to be constructed from the existing ID variables in order to match panel cases. This identifier was generated by creating, and then concatenating, two non-unique identifiers: the village ID and the household number. Because no village ID appears in the raw data (only the village name), villages were assigned persistent, unique three-digit codes

[27] The only case where we find evidence that this could constitute a problem is for time collecting water, we address this problem explicitly in the results section.

for matching purposes. The household number is the two-digit household order suffix from the original household ID. This number is not unique in itself (it starts over from 01 for each primary sample unit), but when combined with the village ID, the resulting identifier is almost unique (in the sense that this combined variable uniquely identifies most observations) and persistent between rounds. Some duplicates resulted in the baseline dataset in a few cases when two EAs existed in the same village, causing the household number to repeat; these were manually matched to follow-up households and assigned new IDs.[28] Once all of the remaining duplicate IDs were corrected, all follow-up variables were assigned a "mid_" prefix and the two datasets were merged.

A total of 871 panel cases were successfully merged, equivalent to 27 percent of the households surveyed at baseline. Note that this corresponds to all the households surveyed by the IEMS, which includes not only Phase-A and Phase-C villages, but also villages that take in part in the studies of the health and water urban activities of the MCA-Lesotho Compact. When restricting the dataset to only households living in the Phase-A or Phase-C villages, there were 673 panel cases, equivalent to 71 percent of the A- or C-village households surveyed at baseline.

The treatment/control status for the rural water intervention was assigned based on village. Villages were assigned to one of three groups: Control (Phase C); villages where treatment started before follow-up (Phase-A^{rev}); and villages originally assigned to treatment, but were not treated until after follow-up data collection (Phase A_1). In the baseline dataset, village names used by BoS did not always correspond to those listed in the DRWS group classifications. So, we matched villages to Phases C, A^{rev}, or A_1 based on a combination of their enumeration area codes, GPS coordinates, or village names.

Village names in the follow-up dataset corresponded more closely to the village names in the group classifications. So, households were matched to treatment in the follow-up data set using the district and village name combinations. From there, for the panel households, we cross-checked the follow-up treatment to the baseline treatment to identify any households that moved between treatment groups between baseline and follow-up. There were no such households. 673 panel households correspond to one of the three study groups (C, A^{rev}, or A_1) while the remaining 198 correspond to villages that are not part of this study or that could not be matched to any of the study groups at baseline and follow-up. The reason for which some villages could not be matched was that village names were not standardized, which made it difficult to match between data set rounds. Table 3 presents the number of matched and unmatched households per group.

Table 3: Households in Baseline, Follow-up, and Panel

Village Group	Baseline Only	Follow-up Only	Panel
Treatment (A^{Rev})	107	175	290
Treatment (A_1)	23	6	80
Control (C)	149	250	303
Total	279	431	673

[28] After sorting by geographic location, the merge was conducted using household level data.

673 households in 72 villages were matched between baseline and follow-up datasets – 370 households in the treatment villages, and 303 in the control villages. There were 279 households in treatment or control villages that were present only in baseline and were unable to be matched with the follow-up dataset – 130 in treatment villages and 149 in control households. 431 households in treatment or control villages at follow-up were unable to be matched to a baseline household – 181 in treatment villages, and 250 in control villages.

The above discrepancies between the baseline and follow-up samples arose because either a household could not be matched between baseline and follow-up, or because there were villages surveyed in the baseline data collection that were not surveyed at follow-up, and vice-versa.

The main circumstance that made it difficult to preserve the panel of households over time was that the names of the village, the unit at which treatment status was assigned, were not useful as unique identifiers: there are many common names used for different villages, sometimes in the same district, and some villages have multiple names that do not resemble each other. In retrospect it would have helped to assign every sampled village a permanent and unique ID code to be reused for each round, and integrate GPS from the beginning to make sure interviewers go to the right village regardless of its name. Some of these limitations could have been tackled during the follow-up fieldwork; however, because NORC did not receive extracts of data during the field period, and only received the actual datasets several months after the end of data collection, most of these issues were discovered much later.[29]

Given that the final panel sample is smaller than originally planned, it is important to discuss the potential consequences of this situation. Sample attrition has two main implications. First, a smaller sample reduces the precision of the estimated impacts. This implies that we may find coefficients that are not significant, or only marginal significant, that with the original sample we would had found significant. In Annex D we discuss updates to the power calculations and conclude that, for most outcomes, it is unlikely that this is a major problem.

The second problem is more serious because sample deterioration could be such that treatment and control groups are no longer comparable. Fortunately, as we discuss in more detail below, we do not find major differences at baseline between treatment and control groups in observable characteristics using the final sample panel, which suggests that randomization was not compromised by sample deterioration.

Finally, even if treatment and control groups in the final sample are balanced, sample deterioration could compromise the external validity of the results. It is worth saying that, in any case, this study was not going to produce results that were representative of a large population (like rural areas in Lesotho), because villages were selected for the study purposefully (as opposed to randomly), so the results are 'representative' only of the households in the selected villages. However, the panel sample (and the results derived from it) may not be representative even of the households in the selected villages due to sample deterioration. To address this possibility, in Annex E we conduct two exercises. First, we use Inverse Probability Weights to correct for sample attrition, we show that the results are not sensitive to this correction. Second, we follow Karlan and Valdivia (2011) and construct different sets of bounds for the treatment

[29] To avoid this in the future, NORC recommended not to do T&M contracts with data collection firms, and use of tablets if possible for data collection. Note that NORC did not have a contract with BoS directly. The BoS contract was with MCA and it was a time and materials contract. Payments were not linked to products and product quality.

impacts in order to assess the extent to which sample attrition may be biasing the results; we find that for most of the outcomes the estimated bounds do not change the conclusions derived from our main specifications.

5. EVALUATION DESIGN

5.1 Overview of Evaluation Design

The original evaluation design for the RWSSA, developed under NORC's first contract with MCC, focused on a randomized design, under which NORC, with MCA, planned for a 6-9 month gap between the end of construction and rehabilitation of treatment water projects in 50 Phase-A villages, and the start of water projects in the 50 control villages (Phase C).

Under the original design, all Phase-A villages had a largely similar construction timeline with concurrent start and end dates of construction; thus, it was reasonable to expect that there would be a nine-month (or, at a minimum, a six-month) lag between the end of construction of the 50 Phase-A villages and the start of construction of the 50 Phase-C villages.[30]

Delays in the construction of Phase-A water systems, resulted in 11 treatment villages (denoted as Phase-A_1 villages) undergoing construction concurrently with the Phase-C control villages. This overlap has called into question the validity of the original evaluation design. Furthermore, as Table 1 demonstrates, construction was completed in only 70 percent of Phase-A villages (34 of the 50) nine months before construction commenced in Phase-C control villages in January 2013. For these 34 villages – which are part of Phase A^{rev}– the time of exposure to treatment before controls began receiving treatment ranges from 10 to 19 months. Since the follow-up data collection preceded the start of construction in Phase-C villages (i.e., November to December, 2012), duration of exposure to treatment by the time of follow-up data collection for these 34 Phase-A^{rev} villages was about 8-17 months.[31] The remaining five villages in Phase A^{rev} and all 11 Phase-A_1 villages had not been exposed to treatment (i.e., construction of their water systems had not been completed) by the time of follow-up data collection.

In what follows, we discuss the implications of these construction delays for the evaluation design and our approach to tackle them.

5.2 Impact Evaluation Design and Methodology

The key aspect of this evaluation is that treatment assignment was randomized, so we can assume that baseline characteristics in treatment and control villages are not different, statistically speaking. Balance tables showing that this is the case are presented in Annex F.

In addition to having treatment randomized, this study also exploits the longitudinal nature of the available data. Specifically, we use a household fixed effects model, which is very similar (and

[30] The requisite time lag to detect effects is measured from the end of construction in treatment group to start of construction in control group (rather than end of construction in control group) because the control conditions change even with the inception of construction/rehabilitation. For example, during the construction period, water supply is interrupted, VIPs in some household are completed and become operational, and the perceptions and attitudes of household members are affected. The end-of-construction in treatment to start-of-construction in control group time frame allows us to avoid such contamination and preserve a largely untouched control group.

[31] Five villages in Phase A^{rev} had no exposure to treatment at follow-up data collection, because completion of construction coincided with or occurred after the follow-up data collection.

in this context, equivalent for most purposes) to the difference-in-differences model. Mathematically, we estimate:

$$Y_{it} = \beta_0 + \beta_1 R_t + \beta_2 I_{it} + \boldsymbol{X}_{\boldsymbol{it}}^{'}\boldsymbol{\delta} + \{\boldsymbol{H_i}\} + \varepsilon_{it} \quad (1)$$

where Y_{it} is an outcome measure of household *i* in round *t*; I_i indicates if the household's village is in the treatment group at follow-up (I_{it} = 1, if household *i* is from a treatment village and *t*=1, and I_{it} = 0 otherwise); $\boldsymbol{X}_{\boldsymbol{it}}^{'}$ is a vector of time-varying characteristics (such as household size age, and education of the head of household); R_t is the round dummy (R_t = 0 at baseline, *t*=0 and R_t = 1 at follow-up, *t*=1), $\{\boldsymbol{H_i}\}$ is a vector of (absorbed) household fixed effects $\boldsymbol{\varepsilon_{it}}$ is an error term and the β_j and $\boldsymbol{\delta}$ are parameters to be estimated.[32] The estimated value of β_2 captures the effect of the program.

As explained above, not all villages in the treatment group were treated in time due to delays in construction work. Specifically, for 11 villages in three districts, construction works had not started when follow-up data was collected, and for 5 villages in two districts construction was not completed when follow-up data was collected. One alternative to tackle this problem is to disregard the original treatment assignment, which we call the "Original Design", and estimate Equation (1) replacing *I* with a dummy variable *I** that is equal to 1 if construction works had ended before follow-up data collection and 0 otherwise (I_{it}^{*}=1, if household *i* is from a treatment village where construction had ended before follow-up data collection and *t*=1, and I_{it}^{*} = 0 otherwise), which we call the "Observed Design". One threat to the internal validity of this approach is that households in villages where construction ended before follow-up data collection may be different than households where construction was delayed, so any result we find at follow-up would confound the treatment effect with differences in these two groups that would had occurred even in the absence of treatment.

The problem is that receiving treatment before follow-up data collection is not only a consequence of being in a village randomly selected into treatment, but also a consequence of other factors. In effect, the 11 villages where construction works started after follow-up data was collected are in districts that are relatively more remote than the other districts; hence, there may also be other factors explaining these delays that can affect the outcomes of interest.

Instead of simply running Equation (1) using *I** as the treatment variable, we can use *I* as an *instrumental variable* of *I**. The objective of this approach is to purge *I** from any factors that determined having received treatment other than randomization. Therefore, the variation of *I** used to estimate the impact of the program only comes from randomization, not from other factors that could bias the estimate of the treatment effect. Mathematically, we estimate a first-stage equation where receiving treatment prior to follow-up data collection (*I**) is a function of *I* and other observable characteristics as

$$I_{it}^{*} = \alpha_0 + \alpha_1 R_t + \alpha_2 I_{it} + \boldsymbol{X}_{\boldsymbol{it}}^{'}\boldsymbol{\vartheta} + \{\boldsymbol{H_i}\} + u_{it} \quad (2)$$

[32] Note that (1) the objective behind including village and round interaction terms is to try to capture any major contextual change at the village level that could affect the outcomes of interest and (2) by including household fixed effects any characteristics that may confound the treatment effect are isolated (controlled), as long as the characteristics are time-invariant..

and then use this model's prediction of I_{it}^{*} (denoted as $\widehat{I_{it}^{*}}$) as the covariate of interest in outcome Equation (1). Concretely, we estimate:

$$Y_{it} = \beta_0^{\mathrm{IV}} + \beta_1^{IV} R_t + \beta_2^{\mathrm{IV}} \widehat{I_{it}^{*}} + \boldsymbol{X_{it}'}\boldsymbol{\delta}^{\mathbf{IV}} + \{\boldsymbol{H_i}\} + \tilde{\varepsilon}_{it} \quad (3)$$

where the superscript IV indicates that these are instrumental variable estimates. The key feature of this technique is that $\widehat{I_{it}^{*}}$ is a function of *I*, the random assignment variable (and other observable characteristics also included in the outcome question), so it is purged of unobservable factors that could confound the effect of the program.

Finally, as an additional specification to estimate the treatment effect we exploit the block design of the program. The fact that treatment was randomized within districts implies that there are treatment and control villages in each district, and that these villages are observationally equivalent within each district. Along these lines, and ignoring sample size issues for a moment, in each district a regression could be run using *I* as the covariate of interest and it would be internally valid because villages in treatment and control groups were randomly selected. To tackle the selection issue created by the fact the some villages in the treatment group were not treated, we can simply run the regressions using the districts where construction works did end in time, dropping the rest of the districts from the analysis. We call this method 'Matching', in the sense that we are restricting the sample to the districts where there are both treatment and control villages, and dropped the districts were, due to construction delays, there are less villages actually treated than what was expected.

Clearly this creates a sample size problem because we would be dropping as many as five out of eleven districts. Furthermore, this method has no external validity, not even in the context of the 100 villages in the original design, because it is not documenting what the program effect would be in districts that are not in the restricted sample. Nevertheless, this approach does have internal validity and together with the specifications presented before, provides a more comprehensive outlook of the program impact.

So far we have focused on the construction of standpipes as the key treatment component of the program; however, construction of VIP latrines was also a part of the program and this may have not coincided with the completion of standpipe construction. To measure whether households have a VIP latrine we use a question on IEMS asking for the type of toilet the household uses, with one of the response options being a VIP latrine. To evaluate this component of the program we can use the same approach presented in Equation (2), but instead of *I** as the treatment variable we use the dummy variable *L* that is 1 if households reported having a VIP latrine and 0 otherwise.

However, because we only have one instrument, the random treatment assignment (*I*), we cannot estimate the "structural" or independent effects of the two components of the program separately because the random variation that we are exploiting comes from the same variable. Whether this is a problem or not may depend on the outcome we are looking at. For example, if we want to evaluate the effect of the program on access to an improved water source, we can discard the possibility that VIP latrine construction had any impact and focus on the water-system construction as the treatment. On the other hand, if we want to analyze the program's effect on toilet use, then we should focus on VIP latrines.

Unfortunately, for the rest of the outcomes it is less clear which component of the program can be discarded. Diarrhea incidence or medical expenditures, for example, may be affected by both improved-water access and VIP latrines. In this case, a decision needs to be made on defining what the treatment is – and, in particular, for the purpose of implementing the instrumental variable approach. Given the importance of access to clean water to prevent infectious diseases, for all outcomes (except toilet use) we focus on construction of water systems as the treatment and for the instrumental variables specification.

6. RESULTS

In this section we present the main impact evaluation results. We first present descriptive statistics for Short-term, Intermediate and Long-term outcomes, and then move to the discussion of the main results.

6.1 Descriptive Statistics

The data used in this evaluation are the baseline and follow-up Impact Evaluation Multipurpose Survey (IEMS). The final panel sample corresponds to 673 households. Of these households, 370 correspond to the original treatment group (A), and 303 to the control group (C).

Short-term outcomes

Table 4 shows descriptive statistics at baseline pooling the two experimental groups, and at follow-up discriminating between the two groups. The results are divided into short-term and intermediate outcomes. We bundle treatment and control groups at baseline for ease of exposition and because these variables are very well balanced at baseline (see Annex F).

Table 4. IEMS Summary Statistics - Short-term outcomes

			Follow-up			
	Baseline		Control		Treatment	
	Mean	N	Mean	N	Mean	N
HH has improved water source	0.58	673	0.15	303	0.55	370
Percent of HH members using toilet	0.36	673	0.31	299	0.65	367
Toilet used by all HH members	0.24	673	0.19	303	0.38	370
Time spent collecting water per day (all sources)	105	653	100	221	58	248
Time spent collecting water per day (main source)	82	653	93	221	49	245

Source: Baseline and Follow-up IEMS Surveys. Samples vary due to item-specific missing data.

At baseline, 58 percent of the households had as their main water source an improved water source, which means having access to sources like a public standpipe or a protected spring, as opposed to an unimproved source, which could be an unprotected spring or surface water.[33]At follow-up the results are somewhat surprising. In the control group only 15 percent of households had an improved water source as their main source, while for the treatment group the figure is 55 percent. This indicates that while the control group experienced a decline in their access to improved water sources, the treatment group barely kept the same level of access to improved water sources. This could be a consequence of the severe drought that hit Lesotho in 2011-2012. That being said, it could be argued that at the very least the RWSSA project helped attenuate the negative impact of the drought on access to improved water sources in treatment villages.

[33] For a complete description of how this indicator is constructed see Annex J and the baseline and follow-up instruments.

The observed changes in toilet use suggest a positive impact of the RWSSA. At baseline 36 percent of the household members used a toilet, while 24 percent of the households report that *all* their members use a toilet. These figures are slightly lower for the control group at follow-up and considerably higher for the treatment group at follow-up. In fact, for the treatment group 65 percent of households members use a toilet, and 38 percent of the households report that all their members use a toilet.

The RWSSA also had positive impacts on time spent collecting water. At baseline households spent, on average, 105 minutes per day collecting water (considering all sources); at follow-up this figure practically remains unchanged for control households (100 minutes per day), but for treatment households falls to 58 minutes per day, a reduction of almost half. A similar analysis can be drawn from the figures for time collecting water only from the main source.

Intermediate outcomes

The results for the Intermediate Outcomes, displayed in Table 5, portray a less-clear picture in terms of the effects of the program. Ten percent of households reported having a member with diarrhea in the past 2 weeks at baseline. At follow-up the figures for treatment and control groups are slightly higher, 12 percent for the control group and 11 percent for the treatment group. Results by age follow a similar pattern. Eight percent of households reported at baseline that at least one household member 5 years old or older had diarrhea; at follow-up these figures are 8 and 6 percent for the control and treatment group, respectively. Diarrhea incidence increased between baseline and follow-up for children younger than 5 years old.[34] In effect, while 7 percent of the households reported that one households member younger than 5 years old suffered diarrhea in the past two weeks, the rates at follow-up for the control was 15 percent and for the treatment group 12 percent. It is possible that the decline in access to protected water sources may explain this increase in the incidence of diarrhea among the youngest household members; note that this problem was observed more pervasively in control than in treatment households.

[34] We found some data issues with the variable that was design to measure the number of household members younger than 5 years old that had experience diarrhea at follow-up. First, diarrhea incidence using this variable was way higher at follow-up than baseline; in effect, while the incidence at baseline was 7 percent, at follow-up the rates were 45 percent for the treatment group and 53 percent for the control group. Second, we found several cases where households supposedly reported having a child younger than 5 years old, but with no actual children in this age range, according to the household roster. Also, many of the questions subsequent to the diarrhea incidence one, that respondents were supposed to answer if they had answered that they had a child that suffered diarrhea, like what symptoms the child had and whether medical attention was sought, were not answered in several cases. For these reasons, we decided to construct a different variable to measure diarrhea incidence. We used the question that asks whether the respondent sought treatment for diarrhea for any household member under 5 years old, and defined a dummy variable for whether there was an answer to this question. In principle, this question should be answered by all respondents reporting any household member younger than five years old with diarrhea in the household in the last two weeks. For consistency, we used this same question for household members five years old and older to calculate their diarrhea incidence. Note that is not the same as saying that only households that sought treatment suffered diarrhea.

Table 5. IEMS Summary Statistics - Intermediate outcomes

	Baseline		Follow-up			
			Control		Treatment	
	Mean	N	Mean	N	Mean	N
Any HH member had diarrhea (past 2 weeks)	0.10	673	0.12	303	0.11	370
Any HH member (5 or older) had diarrhea (past 2 weeks)	0.08	673	0.08	303	0.06	370
Any HH member (below 5) had diarrhea (past 2 weeks)	0.07	272	0.15	131	0.12	178
Over-5 had >1 incidences of diarrhea (past 2 weeks)	0.03	673	0.04	303	0.01	370
Under 5 had >1 incidences of diarrhea (past 2 weeks)	0.03	270	0.07	129	0.06	178
HH spent money on medical visit (incl. travel)	0.004	673	0.02	303	0.02	370
Over-5 member missed work/study in last two weeks	0.03	673	0.02	303	0.01	370

Source: Baseline and Follow-up IEMS Surveys.

Among the households that reported having a member suffer from diarrhea in the previous two weeks, a few additional questions were asked associated to this situation. These questions gather information on diarrhea intensity (number of occurrences), whether they sought any medical help, and if they missed school or work due to this condition.

Three percent of the households reported that at least one household member 5 years or older suffered diarrhea more than once during the two weeks prior to baseline. At follow-up these figures were 4 percent for the control group and 1 percent for the treatment group. For children younger than 5 years old we observe again an overall increase from baseline to follow-up of this indicator, as 3 percent of the households reported at baseline having a member younger than 5 years old with more than 1 diarrhea episode in the past two weeks, while at follow-up the figure for the control group was 7 percent and for the treatment group 6 percent. Very few households spent any money on medical visits. At baseline only 0.4 percent spent any money on medical visits, and at follow-up the figure is 2 percent for both experimental groups. The fraction of households members five years old or older missing work or school in the past two weeks due to diarrhea was 3 percent at baseline and, at follow-up, 2 percent in the control group and 1 percent in the treatment group.[35]

In sum, there do not seem to be significant impacts of the program on Intermediate Outcomes though, perhaps, it would be more precise to say that there have not been any effects on diarrhea incidence. While in all cases households in the treatment group have lower incidence rates at follow-up with respect to the control group, the differences are rather small.

Long-term impacts

Table 6 presents summary statistics for the long-term outcome indicators, which are labor outcome indicators for the most part. Only one indicator had data for both baseline and follow-up, namely, whether or not a household used time saved from water collection to go to work.[36] A

[35] Note that the denominator for the fraction of households spending money on medical visits due to diarrhea and missing work/study due to diarrhea are the total number of households, not only the very few that observed any diarrhea episodes.

[36] Baseline survey does have data on number of *days* that each household member worked in the past two weeks, so in principle this could be used at least to document whether households members worked at all in the past two weeks; however, without knowing exactly how respondents would answer this question when they worked less than one day, we refrained from using this information in this analysis.

small percentage of households used time saved from water collection for work – only 2% of the baseline sample, and only one household out of 673 in the follow-up sample.

Table 6. IEMS Summary Statistics for Long-Term Outcomes

	Baseline		Follow-up			
			Control		Treatment	
	Mean	N	Mean	N	Mean	N
Used time saved from water collection for work	0.02	673	0.00	303	0.00	370
Number of HH members who worked at least 1 hour (past 2 weeks)	N/A	N/A	0.75	303	0.83	370
Total Hours worked in the past 2 weeks for money (men)	N/A	N/A	42.0	283	38.3	349
Total Hours worked in the past 2 weeks for money (women)	N/A	N/A	22.7	285	27.4	351
Total Hours worked in the past 2 weeks for money (per capita)	N/A	N/A	18.3	272	18.5	334
Total Hours worked in the past 2 weeks for money by all members 13+	N/A	N/A	64.3	272	66.5	334
Any man in HH worked for one hour or more in the past 2 weeks	N/A	N/A	0.38	283	0.37	349
Any woman in HH worked for one hour or more in the past 2 weeks	N/A	N/A	0.24	285	0.32	351
Any HH member older than 13 worked for one hour or more in the past 2 weeks	N/A	N/A	0.51	272	0.56	334
HH experienced improved income in last month	N/A	N/A	0.11	303	0.10	370
Total cash income from all sources in 2012	N/A	N/A	8,266	259	11,791	324

Source: Baseline and Follow-up IEMS Surveys. Samples vary due to item-specific missing data.

With respect to the indicators for which only follow-up data is available, in most cases the differences between treatment and control groups do not seem very large, although they reflect an interesting gender-pattern, as households in the treatment group experienced higher female participation in the labor market, relative to households in the control group. In effect, the number of total hours women worked in the past two weeks is 27.4 in the treatment group and 22.7 in the control group, while, for men the results are 38.3 in the treatment group and 42 in the control group. It can also be seen that the fraction of households were at least one women worked in the past two weeks is higher for the treatment than for the control group (32 percent for the treatment group and 24 percent for the control group) while for men the difference is negligible (37 percent for the treatment group and 38 percent for the control group). In other words, females in the treatment households experienced better labor outcomes than their counterparts in the control group, but such pattern was not observed by males.

Finally, the difference in the fraction of households that reported having experienced an income improvement in the last month is basically the same between the treatment and control groups; while for total cash income, we found that the treatment group earned, on average, 3,525 LSL more than did the control group, which is a substantial difference relative to control mean (43 percent).

6.2 Program Impacts

In this section we discuss the regression results following the models presented in Section 5.2. Table 7 presents estimations for the short-term outcomes. In the first row the results for whether the household has an improved water source are displayed. Four different models are presented for this and most of the rest of the outcomes of interest. In the first column the results for the Original Design are presented. In this case the treatment parameter corresponds to a dummy variable for being in the original treatment group, Phase A, regardless of when actual construction works were completed. We can see that the program has a positive and significant impact, the parameter implies that a household being assigned to the original treatment group increased the likelihood of using an improved water source by 34 percentage points, compared to the control group. In Column 2 the displayed parameter corresponds to a dummy for the villages where construction works ended before follow-up data collection. In this case, called the Observed Design, we can see that the treatment effect is slightly higher than what was estimated for the Original Design.

Note that these two first estimates may be biased for different reasons. First, if the program indeed has an effect on this outcome, the coefficient on the Original Design may be an underestimation because the method treats all villages that were originally assigned as if they were treated, though not all were. Second, when we compare only villages that actually were treated with the rest of the villages (Observed Design), we may also get a biased estimate of the treatment effect because households in the subgroup of villages in the original treatment group that did get treated may be have been different from households in the subgroup of villages in the original treatment group that were not treated. If, for example, villages that were actually treated were more likely to have access to improved water source in the absence of treatment than households that were not treated (as it may well be the case, given that some of the villages were construction were delayed were located in more remote places), then the Observed Design parameter may overestimate the causal effect of the program because it would confound the program effect with differences in initial access to improved water source (i.e., that treated households were more likely to have an improved water source anyway).

This is why, in the third column, we present the results for the Instrumental Variable Method, which is our preferred specification. This approach uses the randomness of the Original Design to control for the endogeneity in the Observed Design. In this case, the estimated effect is 50 percentage points (for 1st stage results see Annex G). Lastly, in the fourth specification we simply drop the five districts where construction works were not finished before follow-up data collection. In this case we find that the estimated parameter implies that households in the treatment group were 44 percentage points more likely to have an improved water source at follow-up than the control households.

In the second row, the results for time spent collecting water from all sources are displayed. The coefficients across all four specifications are negative, although they are only significant for the IV approach. The IV coefficient implies that households in the treatment group saved 43.8 minutes per day in time collecting water compared to the control group, which is roughly half the baseline mean. Similar results are observed for time collecting water from the main source. However, for this outcome, the estimate for our preferred specification, the IV model, is only marginally significant, indicating that the effect of the program on time savings from collecting

water from the main source is smaller (and not significant) than the effect associated with all sources of water.

Table 7. The effect of RWSSA on Short-term Outcomes

Outcome	Original Design	Observed Design	Instrumental Variable	Matching
HH has improved water source	0.34***	0.39***	0.50***	0.44***
	(0.089)	(0.093)	(0.070)	(0.100)
	[1346]	[1346]	[1346]	[850]
Time spent collecting water per day (all sources)	-30.7	-32.1	-43.8*	-20.5
	(16.8)	(16.4)	(17.7)	(20.5)
	[1111]	[1111]	[1111]	[710]
Time spent collecting water per day (main source)	-17.5	-38.0**	-25.2	-11.6
	(14.2)	(13.3)	(13.5)	(19.2)
	[1110]	[1110]	[1110]	[708]
Percent of HH members using toilet [(a)]	0.29***	0.47***	0.59***	N/A
	(0.048)	(0.043)	(0.055)	
	[1339]	[1339]	[1339]	
Toilet used by all HH members [(a)]	0.17***	0.25***	0.35***	N/A
	(0.042)	(0.040)	(0.062)	
	[1346]	[1339]	[1339]	

(a) Treatment variable for the 'Observed Design' and IV estimations: Owning a VIP latrine

Notes: Standard errors clustered at the village level in parenthesis, except for the IV where the standard errors are bootstrapped. Sample sizes are in brackets. All the models include household fixed effects and the following covariates: the number of household members; number of household members under 5; number of elderly household members; sex of household head; age of household head; and dummy variables for the education level of the household head. A few outliers are dropped from the time collecting water regressions, in particular 11 observations are dropped from the time collecting water from all sources, and 9 from the time collecting water from main source. These observations were dropped because reported time collecting water exceeded 8 hours a day. We analyzed the results also if we dropped cases where time collecting water exceeded 3 hours a day and the results did not change substantially.

F-statistic for the 1st stage of the IV are 656 when the treatment is construction work and 178 when the treatment is having a VIP latrine.

Source: Baseline and Follow-up IEMS Surveys

* p<0.05 ** p<0.01 *** p<0.001

Note that the sample size for the two variables related to time collecting water is much smaller than for the other outcomes in the table due to item-specific missing data. We explored whether the reason for this were changes in the structure of the surveys between baseline and follow-up, but did not find evidence that these changes were a major cause for the missing data problem.[37] To evaluate whether this missing data problem has any implications on the presented regression results, for time collecting water we model the missing data process and weight the regressions using Inverse Probability Weights. We found that the results are not sensitive to the use of these weights (for details see Annex H).

The last two short-term outcomes are about toilet use. As was explained in Section 5.2, in this case the treatment variable that we use for the Observed Design and IV approach is not a dummy variable for whether or not construction works had ended before follow-up data collection, but a dummy for having a VIP latrine according to the IEMS data. Also, because the relevant treatment for this outcome was having a VIP latrine, there is no point in restricting the sample to the districts where construction works ended as planned, so for these two outcomes we do not present the analysis using the Matching approach.

There are positive and significant impacts across all specifications for both the percent of household members that use a toilet, and the dummy for whether all household members use a toilet. In our preferred specification, the IV approach, we found that the program has increased the average percentage of household members that use a toilet by 59 percentage points, and the likelihood that all household members use a toilet by 35 percentage points.

Table 8 presents regression results for Intermediate Outcomes. Given the analysis of the descriptive statistics presented in Section 6.1, it is not surprising that no coefficient is significant at 5 percent. It is worth highlighting that, however, in most specifications the parameters have the expected negative sign. That is, being in the treatment group seems negatively correlated with diarrhea incidence and the costs associated with it (medical expenditures and loss of days at work/school). For example, for the outcome of having any household member with diarrhea in the past two weeks, the effects are negative for all specifications except for the Observed Design (suggesting positive selection). For the IV and the Matching specification the effect is almost 3 percent, which is large relative to the baseline value of 10 percent. Along these lines, the lack of significant results may be due to the fact that diarrhea incidence was low at baseline anyway, making it hard for any program to have a sizeable effect that can be recovered by the impact evaluation. This is even more critical for other outcomes like whether the households spent any money on medical visits (at baseline less than 1 percent did) or whether days of work or school were missed due to diarrhea (less than 2 percent reported at baseline incurring this type of cost).

[37] In effect, there were logical skips in the follow-up survey that were not present in the baseline survey. These additional skips led to some respondents in the follow-up survey to not have to answer the questions on this variable. Though this effect was not large. See Annex H for more detailed analysis of this issue.

Table 8. The effect of RWSSA on Intermediate Outcomes

Outcome	Original design	Observed design	Instrumental Variable	Matching
Any HH member had diarrhea (past 2 weeks)	-0.020	0.023	-0.028	-0.029
	(0.036)	(0.037)	(0.043)	(0.047)
	[1346]	[1346]	[1346]	[850]
Any HH member (5 or older) had diarrhea (past 2 weeks)	-0.010	0.031	-0.015	-0.021
	(0.028)	(0.026)	(0.039)	(0.033)
	[1346]	[1346]	[1346]	[850]
Any HH member (below 5) had diarrhea (past 2 weeks)	-0.076	-0.032	-0.12	-0.073
	(0.064)	(0.074)	(0.094)	(0.094)
	[581]	[581]	[581]	[360]
Over-5 had >1 incidences of diarrhea (past 2 weeks)	-0.029	-0.0075	-0.042	-0.031
	(0.017)	(0.017)	(0.026)	(0.018)
	[1346]	[1346]	[1346]	[850]
Under-5 had >1 incidences of diarrhea (past 2 weeks)	-0.042	0.055	-0.068	-0.036
	(0.038)	(0.039)	(0.074)	(0.038)
	[577]	[577]	[577]	[357]
HH spent money on medical visit (incl. travel)	-0.012	-0.0072	-0.017	-0.021
	(0.012)	(0.012)	(0.018)	(0.018)
	[1346]	[1346]	[1346]	[850]
Over-5 member missed work/study in last two weeks	-0.0048	0.011	-0.0070	-0.019
	(0.021)	(0.020)	(0.027)	(0.019)
	[1346]	[1346]	[1346]	[850]

Notes: Standard errors clustered at the village level in parenthesis, except for the IV where the standard errors are bootstrapped. Sample sizes are in brackets. All the models include household fixed effects and the following covariates: the number of household members; number of household members under 5; number of elderly household members; sex of household head; age of household head; and dummy variables for the education level of the household head. F-statistic for the 1st stage of the IV is 656.

Source: Baseline and Follow-up IEMS Surveys

* p<0.05 ** p<0.01 *** p<0.001

The only outcome where the estimated coefficients come close to being significant is having a household member five years old or older with more than one diarrhea episode in the past two weeks. In this case the IV estimate shows a decline of 4.2 percentage points and the Matching estimate shows a decline of 3.1 percentage points, practically eliminating all likelihood that a household in the treatment group would suffer this problem (the mean for the control group at follow-up is 4 percent). While not significant at 5 percent, the p-values for these two estimates are both 10 percent.

Finally, Table 9 presents results for long-term outcomes. No coefficient is significant with the exception of the indicator variable for whether at least one woman worked in the past two weeks, and the parameter is significant only for the IV specification.

Table 9. The effect of RSW on Long-Term Outcomes

Outcome	Original Design	Observed Design	Instrumental Variable	Matching
Used time saved from water collection for work	-0.018	-0.0030	-0.026	-0.0020
	(0.014)	(0.014)	(0.018)	(0.013)
	[1346]	[1346]	[1346]	[850]
Number of HH members who worked at least 1 hour in the past 2 weeks.[(a)]	0.043	0.054	0.063	-0.033
	(0.091)	(0.093)	(0.10)	(0.12)
	[673]	[673]	[673]	[425]
Total Hours worked in the past 2 weeks for money (men)[(a)]	-7.50	3.56	-10.8	-10.7
	(8.29)	(8.21)	(8.34)	(11.1)
	[632]	[632]	[632]	[408]
Total Hours worked in the past 2 weeks for money (women)[(a)]	4.02	5.49	5.81	1.89
	(4.91)	(4.71)	(5.84)	(7.23)
	[636]	[636]	[636]	[403]
Total Hours worked in the past 2 weeks for money (per capita)[(a)]	-0.86	2.02	-1.21	-3.05
	(4.24)	(4.15)	(4.60)	(5.73)
	[606]	[606]	[606]	[391]
Total Hours worked in the past 2 weeks for money by all members 13+[(a)]	-2.14	10.2	-3.03	-5.80
	(11.2)	(11.1)	(10.7)	(15.2)
	[606]	[606]	[606]	[391]
Any man in HH worked for one hour or more in the past 2 weeks [(a)]	-0.041	-0.025	-0.059	-0.067
	(0.042)	(0.043)	(0.053)	(0.050)
	[632]	[632]	[632]	[408]
Any woman in HH worked for one hour or more in the past 2 weeks [(a)]	0.069	0.026	0.10*	0.044
	(0.043)	(0.045)	(0.050)	(0.056)
	[636]	[636]	[636]	[403]
Any HH member older than 13 worked for one hour or more in the past 2 weeks [(a)]	0.011	-0.024	0.015	-0.018
	(0.053)	(0.052)	(0.056)	(0.062)
	[606]	[606]	[606]	[391]
HH experienced improved income in last month[(a)]	-0.0033	0.015	-0.0048	0.014
	(0.025)	(0.023)	(0.035)	(0.029)
	[673]	[673]	[673]	[425]
Total cash income from all sources in 2012[(a)]	3801.9	6636.0	5587.2	5320.4
	(3941.0)	(5310.8)	(5396.4)	(6262.4)
	[583]	[583]	[583]	[363]

Notes: Standard errors clustered at the village level in parenthesis, except for the IV where the standard errors are bootstrapped. Sample sizes are in brackets. All the models include household fixed effects and the following covariates: the number of household members; number of household members under 5; number of elderly household members; sex of household head; age of household head; and dummy variables for the education level of the household head. F-statistic for the 1st stage of the IV are 656 when the treatment is construction work and 178 when the treatment is having a VIP latrine.

[(a)] Data only available at follow-up. Does not include household fixed effects. Standard errors are clustered at the village level, except for the instrumental variable estimations.

Source: Baseline and Follow-up IEMS Surveys

* p<0.05 ** p<0.01 *** p<0.001

Given that the (short-term) results presented before showed evidence that the program has reduced the amount of time household members spent collecting water, perhaps it is puzzling that no effects are found on labor outcomes. This is particularly true for female labor outcomes, as time savings from collecting water mostly affect women given that, according to the baseline data, females are in charge of collecting water in almost 80 percent of the households.

It is possible that time availability does not translate into better labor outcomes because these outcomes are restricted not by time availability but by other conditions, like labor demand. It is possible that local labor markets cannot absorb much more labor supply, especially when we consider that time savings probably were observed by most people in each treatment village, rather than just the surveyed households. This type of general equilibrium effect should be addressed in future research.

Another possible explanation for not observing changes in labor outcomes is that for this particular outcome, we are underpowered in part as a consequence of the sample deterioration discussed in previous sections of this report. As we show in Annex D, the impact that the program needed to have on hours worked for women for the current sample to detect treatment effects with acceptable precision was relatively high. Therefore, we cannot discard the possibility that, with a bigger sample, we might have been able to estimate treatment effects with acceptable precision. In any case, given the impact on time savings, any impact the program may have had on hours worked for females would probably have been relatively small.

On the other hand, it is important to recognize that while the program reduced the amount of time households members spend collecting water, this reduction was probably not enough for the evaluation to detect a significant effect on hours worked.

This also underscores the importance of the results on whether at least one woman worked in the household, which are significant under the IV approach. Taken together, these results suggest that while time savings may not had been enough to significantly increase the number of hours women participate in the labor market, more women are working at least a few hours. Further research should address whether households that saved the most time collecting water also observed an increase in female participation in the labor market.

In sum, the RWSSA has had substantial short-term effects. The program is associated with greater access to improved water sources and greater toilet use, and less time spent collecting water. In terms of intermediate outcomes, no statistically significant effects are found for diarrhea incidence, although the signs for most of the analyzed variables (and in the case of the IV, all of them) indicate a negative correlation between the program and diarrhea incidence and its costs. Finally, the evaluation found very weak effects on female labor outcomes.

7. CONCLUSIONS

This report presents the evaluation results of the Lesotho Rural Water and Sanitation Project Impact Evaluation. In this project, treatment assignment was randomized, so differences at follow-up in the outcomes of interest could be attributed to the program. We analysed Short-term, Intermediate and Long-term effects.

The main challenge of this evaluation was related to data processing. Specifically, it was particularly time consuming to link the baseline and follow-up household data. The main reasons for this were that village names were not standardized, which made it difficult to match between data set rounds. It is also possible that due to confusion with village names, some villages that were not part of the study were surveyed, in lieu of villages that were part of the study. As a consequence, the final sample of households only covers 72 villages, instead of the planned 100.

The impact evaluation shows that the program has had important effects on some wellbeing indicators. For short-term outcomes we found that households in the treatment group are more likely than in the control group to use as their main water source an improved water source, such as a public standpipe or a protected spring, as opposed to an unimproved source, such as an unprotected spring or surface water. They are also more likely to use a toilet and spend less time collecting water.

The results for the intermediate outcomes are much weaker than those observed for the short-term outcomes. All of the analyzed intermediate outcomes are related to diarrhea incidence and its costs. We did not find any impacts significant at 5 percent; however, all the coefficients, at least for our preferred specification, (the Instrumental Variables) have the expected signs. The coefficients are also negative when the outcomes are whether households spend any money on medical visits due to diarrhea, or whether household members missed any work or school due to diarrhea. It is possible that the estimated effects of the program on diarrhea incidence are not statistically significant because diarrhea had a relatively low incidence rate at baseline to start with. The modest impact on diarrhea may also be because the quality of consumed water may have not improved significantly. Because water was not tested at the point of consumption, we cannot document the extent to which water quality actually improved or not. Another potential explanation for the lack of significant results could be lower power due to the observed sample deterioration. As we show in Annex D, we do not think this was the key issue behind the documented weak results.

Finally, no major impacts on labor outcomes are observed. This is somewhat puzzling as more available time (due to reductions in time collecting water) could have translated into more time working. It is possible that time availability did not translate into better labor outcomes because the latter was not restricted by time availability but by other conditions, like the labor market itself. It is not clear that local economies would have the capacity to absorb the shift in labor supply that a program like this may had caused in treatment villages, at least not automatically. That being said, it is important to highlight that the program increased the likelihood that at least one women would work, although no impacts were found in terms of hours worked by women.

In terms of policy implications, the results described in this report imply that this type of program can have major impacts on households wellbeing via reductions on time spent collecting water, but limited effects on outcomes that may seem more important, like diarrhea

incidence. Furthermore, even if household members spend less time collecting water, it is not clear that this will translate in greater labor force participation, as labor outcomes may depend on more factors than just greater labor supply.

On the other hand, it would be a mistake to undervalue the importance of reducing time collecting water. It is possible that these time savings will have effects on outcomes that cannot be observed by an instrument like the one fielded in the context of this evaluation. For example, more available time for children could have an effect on time studying, which could have an effect on test scores. More time studying and greater academic achievement will presumably translate in greater opportunities for children in the future.

Regarding further research, given that both treatment and control households have been treated for a long time now, it is not likely that repeated collection of household data for the purpose of analyzing the outcomes that were analyzed in 2012 would improve our understanding of the program impacts. However, there are still a number of important research questions that could be addressed and policy lessons learned if a third wave of data were collected. NORC would like MCC to consider two different options for this.

In the first option we propose to answer two sets of research questions. We would analyze the extent to which the water systems, the Water Minders, and Village Water and Health Committees set up under the project are operating as expected – and, if not, why. Research activities under this option would also assess the continued functioning of the VIP Latrines in the households and the degree to which households have retained knowledge and practices of proper hygiene and sanitation as learned through the PHAST trainings. For this option, we propose conducting modified versions of the Activity Monitoring Plan (AMP) surveys, which *inter alia* comprise a VWHC Questionnaire and a WM Questionnaire. In parallel, a household survey would be used to assess the continued functioning of the VIP Latrines and household knowledge and practices of proper hygiene and sanitation. All three questionnaires would provide input into whether the WMs and Village Water and Health Committees are actually doing what they were trained to do, and what exactly has been their *de facto* role in the villages.

The second option focuses on studying the long-term impacts of having access to improved water and sanitation during early childhood. For this we propose fielding a household questionnaire in 2017 comprising a subset of questions contained in the one fielded in 2012, but supplemented with items related to child development, for example, height, weight, and school performance indicators. The population of interest would be 6-8 years old children in treatment and control villages. These children were *in utero* or younger than 2 years old in 2011-2013, when treatment and control group were subjected to variations in program exposure. It is worth highlighting that for this analysis it is precisely the fact that both treatment and control groups have been treated since 2013 what would allow us to look at long-term outcomes of differential exposure during early childhood (or *in utero*). This is because the length of the interval guarantees that any differences in outcomes at endline (in 2017) can be attributed to access to improved water in 2011-2013, not to a more prolonged difference in exposure to treatment. There is a rich literature on the importance of early childhood for human development. From parental smoking to weather shocks to education interventions, the empirical literature shows the importance of a person's early years for later human development. Conducting an endline of RWSSA would allow us to analyze the long-term effects of access to improved water in early childhood, which could constitute an important contribution to this field.

REFERENCES

Akuoko-Asibey and McPherson (1994). Assessing hygiene and health related improvements of a rural water supply and sanitation programme in northern Ghana. *Natural Resources Forum, 18*(1), 49-54.

Andres *et. al.* (2014). Sanitation and Externalities: Evidence from Early Childhood Health in Rural India. *Policy Research Working Papers* (2014)

Aziz *et. al.* (1990). Reduction in diarrheal diseases in children in rural Bangladesh by environmental and behavioral modifications. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 84.3 (1990): 433-38.

Boone et. al. (2011). Household Water Supply Choice and Time Allocated to Water Collection: Evidence from Madagascar. *Journal of Development Studies, 47*(12), 1826-1850.

Cairncross, Sandy, et al. Water, sanitation and hygiene for the prevention of diarrhea. International Journal of Epidemiology 2010;39: 193–205.

Checkley *et. al.* (2004). Effect of water and sanitation on childhood health in a poor Peruvian peri-urban community. *The Lancet* 363.9403 (2004): 112-18.

Cowater International Inc. (2016). Project Management, Construction Supervision and Environmental Management of Rural Water & Sanitation Projects (WS-E-011-09), Project Completion Report.

Crow *et. al.* (2012). Community Organized Household Water Increases Not Only Rural incomes, but Also Men's Work. *World Development* 40.3 (2012): 528-41.

Daniels *et. al.* (1990). A case-control study of the impact of improved sanitation on diarrhea morbidity in Lesotho. *Bulletin of the World Health Organization*; 68(4): 455 – 463.

Estimates on the Use of Water Sources and Sanitation Facilities, Lesotho, updated June 2015. WHO/UNICEF Joint Monitoring Program for Water Supply and Sanitation.

Esrey *et. al.* (2000). Drinking water source, diarrheal morbidity, and child growth in villages with both traditional and improved water supplies in rural Lesotho. *Am J Public Health American Journal of Public Health* 78.11 (1988): 1451-455.

Hoque *et. al.* (1996). Sustainability of a water, sanitation and hygiene education project in rural Bangladesh: a 5-year follow-up. *Bulletin of the World Health Organization*; 74 (4): 431-437.

Huttly et. al. (1990). The Imo State (Nigeria) Drinking Water Supply and Sanitation Project. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 84(2), 309-315.

Ilahi *et. al.* (2000). Public Infrastructure and Private Costs: Water Supply and Time Allocation of Women in Rural Pakistan. *Economic Development and Cultural Change* 49.1 (2000): 45-75.

Ilahi (2001). Children's Work and Schooling: Does Gender Matter? Evidence from the Peru LSMS Panel Data. *Policy Research Working Papers* (2001)

Ilahi (2001). Gender and the allocation of adult time: evidence from the Peru LSMS panel data. *Policy Research Working Papers* (2001)

Karlan, Dean and Martin Valdivia (2011). Teaching entrepreneurship: Impact of business training on microfinance clients and institutions. The Review of Economics and Statistics, 93(2): 510–527

Kiendrebeogo (2012). Access to Improved Water Sources and Rural Productivity: Analytical Framework and Cross-country Evidence. *African Development Review* 24.2 (2012): 153-66.

Kolb et. al. (2008). An integrated method for evaluating community-based safe water programs and an application in rural Mexico. *Health Policy and Planning* 23.6 (2008): 452-64.

Lesotho Country Proposal to the Millennium Challenge Corporation (MCC). July 2006. A Programme for Improvement of Water Supply, Rehabilitation of Health Infrastructure and Promotion of Private Business Development.

Nanan et. al. (2003). Evaluation of a water, sanitation, and hygiene education intervention on diarrhea in northern Pakistan. *Bulletin of the World Health Organization: the International Journal of Public Health*; 81(3): 160-165

NORC at the University of Chicago (2015). Revised Evaluation Design Report (Revised: February 23, 2015)

Nankhuni and Findeis (2004). Natural resource-collection work and children's schooling in Malawi. *Agricultural Economics* 31.2-3 (2004): 123-34.

Lesotho Millennium Development Agency (2014). Audit of the Fund Accountability Statement of the Lesotho Millennium Development Agency for the period January 15, 2014 to March 31, 2014.

Lesotho Millennium Development Agency (2015). Audit of the Fund Accountability Statement of the Lesotho Millennium Development Agency for the period April 1, 2014 to March 31, 2015.

Office of Inspector General (2014). Audit report no. M-000-14-010-N, April 7.

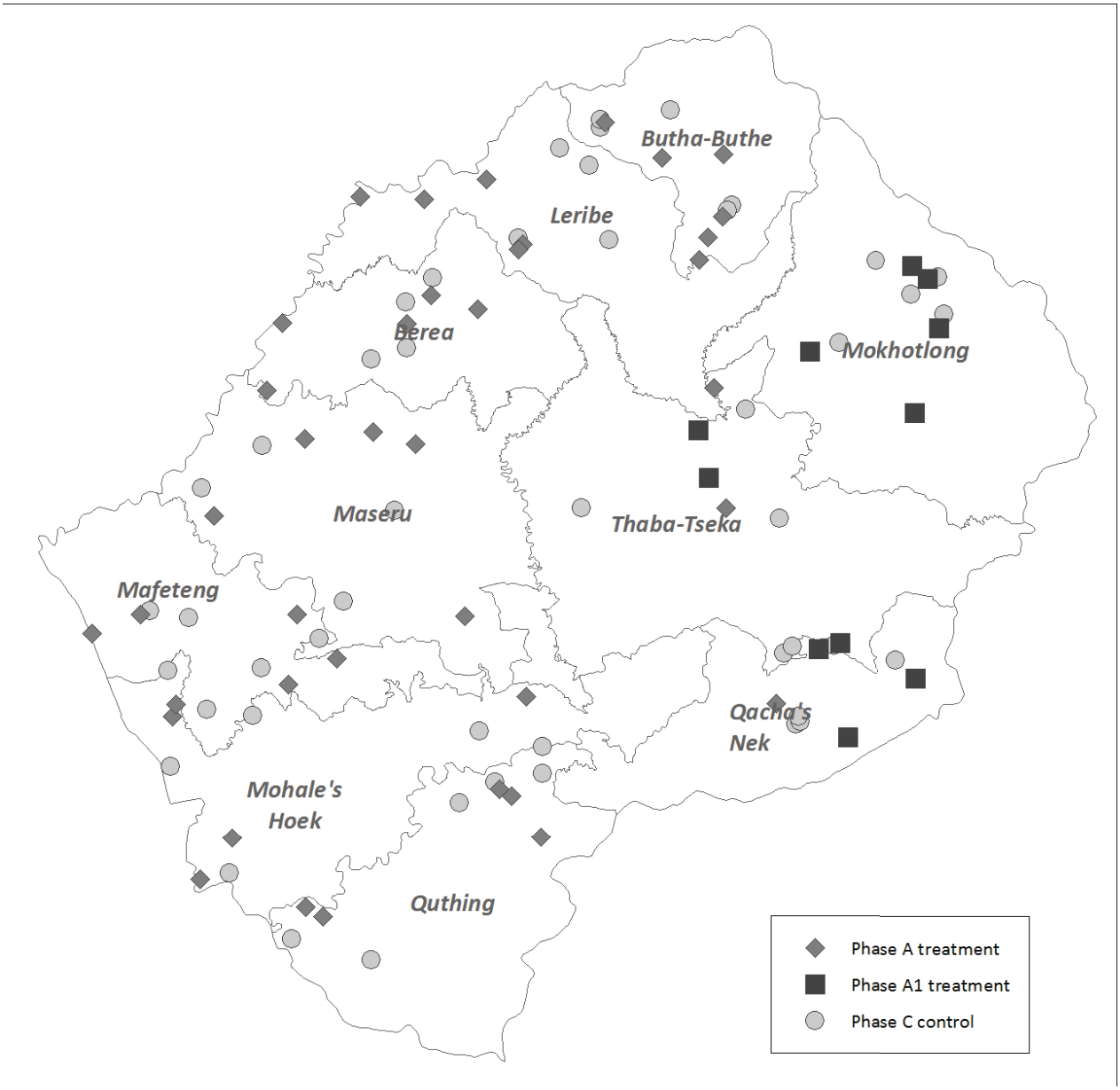
Semba *et. al.* (2011). Relationship of the Presence of a Household-Improved Latrine with Diarrhea and Under-Five Child Mortality in Indonesia. *American Journal of Tropical Medicine and Hygiene* 84.3 (2011): 443-50.

UNICEF/WHO. Progress on Sanitation and Drinking Water – 2015 update and MDG assessment. 2015. Annex 3.

White, Bradley, and White. Drawers of Water: Domestic Water Use in East Africa. 1972. *Bulletin of the World Health Organization*. 2002; 80(1): 63–62.

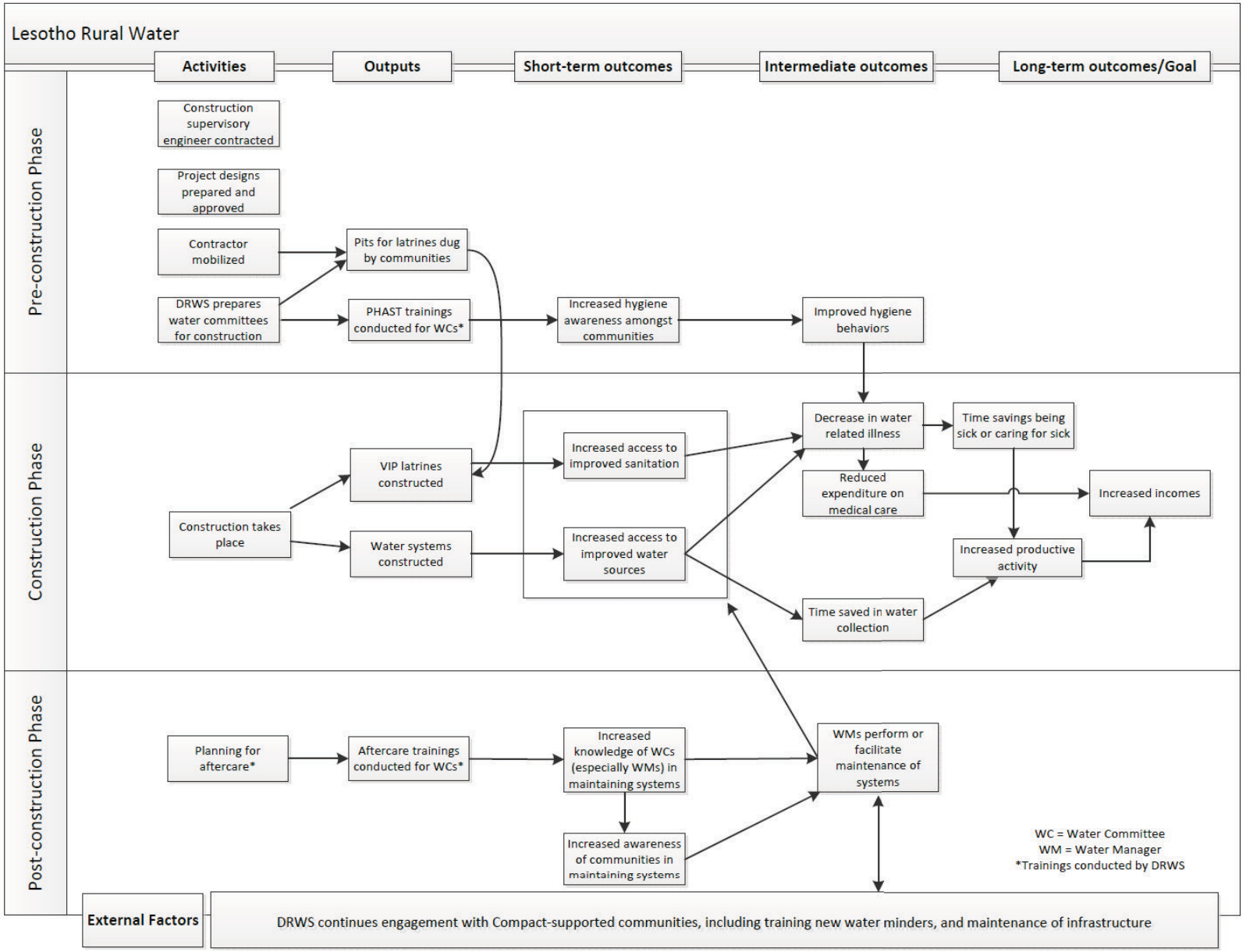
Wilson *et. al.* (1991). Hand-washing reduces diarrhea episodes: a study in Lombok, Indonesia. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 85.6 (1991): 819-21.

ANNEX A: MAP OF LOCATIONS OF TREATMENT AND CONTROL VILLAGES



Note: Some of the geographic coordinates provided by DRWS were inaccurate or insufficiently precise. NORC staff attempted to verify the location of each rural water village using the Lesotho 2006 Census GIS database and external map sources (e.g., Google Earth). While we are highly confident about the locations of the majority of sites, some of the points shown on the map may not represent the exact location of treatment and control sites.

ANNEX B: PROGRAM LOGIC OF LESOTHO RWSSA (UPDATED 2014)



ANNEX C: REVIEW OF LITERATURE PERTINENT TO THE RURAL WATER EVALUATION

Author and title	Intervention	Methodology & data	Relevant conclusions
Effects of increased use of safe drinking water, use of a VIP and better hygiene behavior on incidence of water-related diseases (and other health outcomes).			
WASH (Water, Sanitation, and Hygiene) and Water/Sanitation (no hygiene) Interventions			
Huttly et. al. (1990). The Imo State (Nigeria) Drinking Water Supply and Sanitation Project	Water Supply and Sanitation Project in 3 villages in south-eastern Nigeria	Difference in difference; repeated cross-sectional surveys and longitudinal data ; variables were water diseases (worms and diarrhea, all ages) and nutritional status (children)	No impact on overall period or prevalence of disease in cross sectional data but impact on diarrhea morbidity was found in limited sub-groups. Also significant overall program impact was found on worm disease in longitudinal data. A greater impact of water availability rather than quality was suggested for rates in young children. The prevalence of wasting among children < 3 years decreased significantly with intervention.
Akuoko-Asibey and McPherson (1994). Assessing hygiene and health related improvements of a rural water supply and sanitation programme in northern Ghana	Rural water supply and sanitation programme in northern Ghana	Quasi experimental difference in difference; variables were sources of water collection and storage arrangements, waste disposal, hand washing, food storage arrangements, and defecation and practices associated with feces in program and non-program areas	Some success in water use and water and food storage; little progress in attitudes towards disposal of HH and human waste and in women's knowledge of disease.
Children-specific WASH Interventions			
Aziz *et. al.* (1990). Reduction in diarrheal diseases in children in rural Bangladesh by environmental and behavioral modifications	Water, sanitation and hygiene education intervention in a rural area of Bangladesh	Difference in differences estimator between two similar areas	Children in the program area had 25% fewer episodes of diarrhea than those in the control area. Within the program area, children from households living closer to hand-pumps or where better sanitation habits were practiced experienced lower rates of diarrhea. Suggests that an integrated approach to interventions can have a significant impact on diarrheal morbidity.
Checkley *et. al.* (2004). Effect of water and sanitation on childhood health in a poor Peruvian peri-urban community.	No intervention *per se*; studies a birth cohort of Peruvian children, for health outcomes over 24 months, against baseline data on household water and sanitation conditions.	Difference in difference. Survey data followed up with children once a day for diarrhea and once a month for anthropometry; obtained baseline data on household WASH conditions.	Better water source alone did not accomplish full health benefits. In 24-month-old children from households with a water connection, those without adequate sewage disposal and with small storage containers were 1·8 cm shorter than children in households with sewage and with large storage containers. Children with the worst conditions for water source, storage, and sanitation had 54% more diarrheal episodes than did those with the best conditions.

Author and title	Intervention	Methodology & data	Relevant conclusions
Nanan et. al. (2003). Evaluation of a water, sanitation, and hygiene education intervention on diarrhea in northern Pakistan.	WASEP, a village level project which incorporates engineering solutions with appropriate education to maximize facility usage and improve hygiene practices.	Difference in difference (3 months only), using logistic regression and controls.	Children in control villages had 33% higher odds of diarrhea than children living in program villages. Boys had 25% lower odds of having diarrhea than girls. Relevance: shows how integrated approaches can be successful.
Esrey *et. al.* (2000). Drinking water source, diarrheal morbidity, and child growth in villages with both traditional and improved water supplies in rural Lesotho	Villages in four districts (50 per cent of the rural population of Lesotho) in the lowlands or foothills received an improved water supply between 1967 - 1983, in form of a continually functioning tap or hand pump.	Post-post multivariate regression. Data included morbidity and growth data on 247 children 5 and under in 10 representative villages that had an improved water supply at least one year prior to investigation. Compared those who relied exclusively on improved water to those who relied on it only partially.	Children whose families relied exclusively on the new water supply for their drinking and cooking needs grew 0.438 cm and 235 g more in six months than children whose families supplemented the new water supply with the use of contaminated traditional water for drinking and cooking. Results suggest that improved drinking water supplies can benefit preschool children's health after infancy, but only if they are functioning and utilized exclusively for drinking and cooking purposes
Hoque *et. al.* (1996). Sustainability of a water, sanitation and hygiene education project in rural Bangladesh: a 5-year follow-up	An integrated water supply, sanitation and hygiene education project from 1983-87. Provided hand pumps, pit latrines, and hygiene education to about 800 households. After 1987 no external support was provided to maintain these provisions.	Methodology: Difference in Difference using a cross-sectional data. Data: Baseline and follow-up surveys from 500 randomly selected households from the intervention and control areas.	82% of the pumps still functional 64% of latrines and 84% of the adults were using latrines (vs. 7% in the control area.) The prevalence of diarrheal diseases among the control population was about twice that among those in the intervention area. However, knowledge related to disease transmission, was poor (though, recall that all training had stopped) and similar in both areas. Relevance: importance of continuance of WASH education.
Water-Only Interventions			
White, Bradley, and White (1972). Drawers of Water: Domestic Water Use in East Africa	Improving water volume and quality	Using observational evidence only in 34 sites in 3 countries over 3 years.	Shows positive effect of water quality on water borne diseases; one of the original studies in the field and still considered by some the most comprehensive study to date

Author and title	Intervention	Methodology & data	Relevant conclusions
Kolb et. al. (2008). An integrated method for evaluating community-based safe water programs and an application in rural Mexico	UV Waterworks, a community-based water purification system in rural Mexico	Stepwise evaluation framework of effect on health 5 years after the programme began; variables include physical performance of the water system, community capacity to maintain and manage the systems, and the time and budget constraints of households participating in the program.	No impact on diarrhea incidence was found. (a) household priorities and preferences were a key factor in maintaining exposure to safe drinking water sources, and therefore (b) user convenience was a primary leverage point for programme improvement.
Hygiene-Only Interventions			
Wilson *et. al.* (1991). Handwashing reduces diarrhea episodes: a study in Lombok, Indonesia	Sixty-five mothers were given soap and an explanation of the fecal-oral route of diarrhea transmission, and the message was repeated and reinforced fortnightly	Pre-post evaluation of 65 families	Children of these mothers experienced an 89% reduction in diarrhea episodes
Independently of other interventions, the impact of VIPs on incidence of water borne disease (children only)			
Semba *et. al.* (2011). Relationship of the Presence of a Household-Improved Latrine with Diarrhea and Under-Five Child Mortality in Indonesia	Improved latrines in rural and urban areas of Indonesia	Multivariable logistic regression models	The lack of a household improved latrine is associated with diarrhea and under-five child mortality in Indonesia.
Daniels *et. al.* (1990). A case-control study of the impact of improved sanitation on diarrhea morbidity in Lesotho	Project to improve sanitation in Mohale's Hoek district within Lesotho	Randomized case-control design using clinic-based diarrheal cases and controls that experienced other illnesses. Data collected on child, illness, access to water/sanitation, hygiene practices from interview with caregiver. A random sample visited at their homes and the water / sanitation facilities/general conditions were observed.	Children under-5 from households with a latrine experienced 24% fewer episodes of diarrhea then households without a latrine. The results of this study provide evidence that improved sanitation can have a positive impact in the reduction of diarrhea morbidity in young children in rural Lesotho.

Author and title	Intervention	Methodology & data	Relevant conclusions
Andres *et. al.* (2014). Sanitation and Externalities : Evidence from Early Childhood Health in Rural India	Household level: moving from open to fixed-point defecation OR from unimproved sanitation to improved sanitation. Village level: an external benefit (externality) produced by the neighborhood's access to sanitation infrastructure	Individual production function of health for children assuming a linear-in-parameters approximation, and robust results using several econometric specifications. Main inputs are HH access to sanitation and ratio of access to sanitation at village level. Data: 206,414 children under 48 months in rural areas of India from District Level Household Survey '07-'08.	Finds significant direct benefits and concave positive external effects for both improved sanitation and fixed-point defecation. At HH level, finds 10% reduction in diarrhea from sanitation improvements and 5% reduction from moving from open to fixed point defecation. Combining HH and village interventions finds 47% reduction in diarrhea prevalence between children living in a household *and* a village with improved sanitation vs children without either. 1/4 of benefit is due to the direct benefit rest to external gains.
Effect of improved access to safe drinking water and improved hygiene practices on time-savings to households (collection time, days sick, hours off work)			
Ilahi (2001). Gender and the allocation of adult time: evidence from the Peru LSMS panel data	None; uses panel data from Peru in 1994 and 1997 to examine the impact of water infrastructure on total time spent in housework and in income-generating activities by male and female adults.	Regression estimation using cross sectional data, controlling for unobserved heterogeneity. Data from 1994 and 1997 Peru LSMS panel household data: 898 HH's and 2095 individuals.	Women in households without in-house water supply do not have significantly higher housework burdens than women in households with piped nor do they spend less time in income-generating activities. For men, however, in-house water supply significantly increases time spent in self-employment activities (such as agriculture) and decreases time spent in wage work. Demonstrates potential positive income-generating impacts of piped water supply for men but questions it for women.
Ilahi *et. al.* (2000). Public Infrastructure and Private Costs: Water Supply and Time Allocation of Women in Rural Pakistan	None, uses existing HH surveys to examine access to water-community and household levels and the time allocation of women	Estimation reduced-form time equations. Data from the 1991 Pakistan Integrated Household Survey (2,400 rural household)	Poor water supply induces women to reduce their market-oriented work and thus their contribution to household income (however, impact on overall HH income unclear because male labor response is unclear). Results also indicate poor water supply causes an increase in the total work burden of women and a decrease in their leisure. Improved water supply could change the nature of women's contribution to the household from performing everyday chores to doing income-generating work.

Author and title	Intervention	Methodology & data	Relevant conclusions
Boone et. al. (2011). Household Water Supply Choice and Time Allocated to Water Collection: Evidence from Madagascar	None: uses household survey data from Madagascar	Conditional logit model for household's choice of water source as a function of distance to the source and HH characteristics. Reduced form model regressing collection time on a set of exogenous variables. Data: 2190 household surveys with detailed data on the characteristics of household members, as well as information on household wealth and assets. The employment and time use module included information on time use in various activities in the last seven days, including hours spent gathering water.	Women and girls spend the most time gathering water. However, investments to reduce to the distance to water sources will have larger impacts on adults than children, and on men than women.
Effects of improved access to safe drinking-water and improved hygiene practices on school absenteeism (sick days, school days missed)			
Nankhuni and Findeis (2004). Natural resource-collection work and children's schooling in Malawi.	None-uses existing data and hypothesis only that investigates if long hours of work spent by children in fuel wood and water-collection activities influence the likelihood that a child aged 6–14 attends school.	Two-stage conditional maximum likelihood estimation. Data from the 1997–1998 Malawi Integrated Household Survey (IHS) conducted by the Malawi National Statistics Office (NSO) in conjunction with the International Food Policy Research Institute.	Attendance decreases as hours allocated to resource collection work increase. Having piped water access in the home significantly reduces the probability of and time spent in water collection among children. They also find that having piped water access is positively associated with a child attending school and not doing any water-collection and negatively associated with water collection while attending school. Girls spend more hours on resource-collection work and are more likely to be attending school while burdened by this work and may find it difficult to progress well in school. However, girls are not necessarily less likely to be attending school (may do both collecting and attending). Shows water programs can affect absenteeism and that, furthermore, they can especially have an impact on girls' attendance.
Ilahi (2001). Children's Work and Schooling: Does Gender Matter? Evidence from the Peru LSMS Panel Data	None; uses data from the Peru LSMS Panel Data	Uses reduced-form equations for each individual, controlling for unobserved heterogeneity using the panel properties of data. Data from Peru's LSMS panel on time allocation of boys and girls in both rural and urban areas.	The findings suggest that changes in household water supply affect the schooling and work of girls more than boys. An in-house water supply has a significant impact on grade-for-age of girls but not boys, and seemingly no significant impact on time in housework.
Effects of improved access to safe drinking-water close to home on productivity of economically active members of household (agricultural output, income per hour, days worked)			
See Crow *et al.,* below.			

Author and title	Intervention	Methodology & data	Relevant conclusions
Kiendrebeogo (2012). Access to Improved Water Sources and Rural Productivity: Analytical Framework and Cross-country Evidence	None	Regresses rural labor productivity growth rate on access to drinking water (with controls). Based on sample of 27 African countries over 1990-2010.	The empirical analysis reveals that increasing the access rate to drinking water significantly increases the growth rate of agricultural labor productivity. Surprisingly, the access rate to improved sanitation facilities has not significantly impacted rural productivity growth. However, the positive effect of drinking water access is reinforced by the presence of a better sanitation system.
Effects of time savings gained through improved access to safe drinking-water and reduced incidence of water-related diseases on household income			
Crow *et. al.* (2012). Community Organized Household Water Increases Not Only Rural incomes, but Also Men's Work	None; took a census to compare results of Community-organized, household water improvements in western Kenya and divided villages into 3 arms: (i) unprotected; (ii) protected but not piped; and (iii) protected and piped.	Used mixed methods for (1) with/without comparisons between households obtaining water from different types of springs – unprotected, protected and not piped, and protected and piped; and (2) before/after comparisons for households with protected and piped water. The before/after data is based on respondents' recall. Quantitative survey for prevalence of different water management regimes, and qualitative for characteristics of communities. Data: A spring census was as a sampling frame for the selection of seven villages. Two villages have protected springs and piped homestead connections; two have protected springs but no homestead connection; and three draw water from unprotected springs.	Piped water reduces the work of women and girls, and facilitates home garden and livestock production. Women recognize clear time-benefits. Men, however, experience extra work. Together these changes lead to increased household incomes.

ANNEX D: POWER ESTIMATES

The fact that the final sample size of the panel is smaller than what was originally planned has implications in terms of the minimum effect that can be detected with acceptable precision. As is discussed in the Revised Evaluation Design Report (NORC, 2015), the original evaluation of the program was expected to achieve MDES of between 0.23 and 0.28[38] and 80-percent power, assuming an ICC of between 0.1 and 0.2, a level of significance of 5 percent, and R^2 (covariate capture) of 0.3. Other than the sample size, to update power (or MDES) calculations, we take advantage of the fact that we are now able to include the observed estimated ICC and R^2. Table 10 shows updated MDES calculations for four selected outcomes of interest, given their observed sample distributions and final sample size.

Note that the estimated ICC is pretty high for having access to an improved water source and time spent collecting water (higher than the upper bound used for power calculations in the Revised Evaluation Design Report of 0.2), while for diarrhea incidence is lower than the lower bound; for the number of hours worked for women, the estimated ICC is within the bounds. The fact that ICC is higher for the first two selected outcomes than for the other two is perhaps not too surprising, as poor access to water and distance to water sources may be determined at the village level (in other words, if one household in a small village has poor water access probably all households in that village do), while morbidity or labor outcomes may be much more household-specific.

In the case of the R^2, for the first three selected outcomes we can see that the estimated figures are higher than the 0.3 value assumed in the original calculations. This is because for these outcomes we included household fixed effects in our regressions, so a large fraction of the variability is explained by these fixed effects, increasing the power of the estimates (or reducing the minimum difference that can be detected). For female labor outcomes, on the other hand, no comparable baselined data was available, so the model did not include households fixed effects, reducing the fraction explained by the covariates in the model.

The MDES estimated in the Design Report was lower than the figures observed for the two first two and fourth outcomes in Table 10. However, the effects of the intervention on improved water sources and time spent collecting water were so large anyway that it was still possible to estimate program effects with acceptable precision.

On the other hand, the estimated MDES for diarrhea incidence is within the bounds of the MDES discussed in the Revised Evaluation Design Report. The smaller observed ICC and higher R^2 offset loss of power from the smaller sample, so the MDES for this outcome was within the bounds originally planned. Finally, the observed MDES for number of hours women worked is higher than the upper bound in the Design Report estimates.

[38] MDES is measured in standard-deviation units so these decimal values simply refer to the number (or proportion) of a standard deviation.

Table 10. MDES for selected outcomes based on estimated ICC and actual sample size

Variables	Ave cluster size	ICC	R^2 (covariate capture)	MDES(6)
HH has improved water source	10	0.45	0.63	0.35
Time spent collecting water per day (all sources)	7	0.22	0.61	0.34
Any HH member had diarrhea (past 2 weeks)	10	0.03	0.53	0.25
Total Hours worked in the past 2 weeks for money (Women)	9	0.11	0.10	0.35

Notes: The MDES values are based on a three-stage, blocked, cluster-randomized control trial design. The first stage (block) is the district, the cluster refers to villages, and the tertiary stage is the household. R^2 is the R-squared of a regression containing all covariates but that has been purged of the effect of the treatment. ICC is calculated using (treatment and control) baseline data, except for female labor outcomes, for which no baseline data is available, so the ICC of the control group at baseline is used. The number of districts is 10. The average number of villages per district is 7, and odd number, so to be conservative we present MDES assuming there are 6 villages in each district.

ANNEX E: ADDRESSING SAMPLE REDUCTION

Inverse Probability Weights

To study whether sample attrition had any impacts on the estimated treatment effects, we use Inverse Probability Weights (IPW) to correct for sample attrition and assess the extent to which the coefficients of interest change. To do this, we model the probability that a household surveyed at baseline is not surveyed at follow-up, and then produce IPW weights following:

$$w^{IPW} = 1/(1 - \hat{p})$$

Where $\hat{p}$ is the estimated probability that a household surveyed at baseline is not surveyed at follow-up. These weights overweight households that are more likely to be dropped from the sample, and underweight households that are more likely to be surveyed both at baseline and follow-up, so the weighted sample better resembles the characteristics of the original sample.

We then ran regressions for short-term and intermediate outcomes using these weights. The results of the weighted regressions, shown in Table 11 - **Table 13**, reveal negligible changes compared to those in the main text using unweighted regressions (see Table 7 to Table 9 for the original –unweighted- results), in particular when we focus on the IV results, which is our preferred specification. The only exceptions is for time spent collecting water per day (main source), for which we find significant results when we weight the regressions but not significant when we do not. This suggests that the results presented in the main text of the report are conservative.

Table 11. The effect of RWSSA on Short-term Outcomes Correcting for Sample Attrition

Outcome	Original Design	Observed Design	Instrumental Variable	Matching
HH has improved water source	0.34***	0.39***	0.60***	0.45***
	(0.089)	(0.093)	(0.093)	(0.10)
	[1346]	[1346]	[1346]	[850]
Time spent collecting water per day (all sources)	-31.2	-32.6	-45.0**	-22.0
	(17.1)	(16.7)	(14.7)	(20.8)
	[1111]	[1111]	[1111]	[710]
Time spent collecting water per day (main source)	-17.7	-37.4**	-47.7**	-12.6
	(14.2)	(13.3)	(14.6)	(19.3)
	[1110]	[1110]	[1110]	[708]
Percent of HH members using toilet (a)	0.29***	0.47***	0.67***	N/A
	(0.048)	(0.043)	(0.13)	
	[1339]	[1339]	[1339]	
Toilet used by all HH members (a)	0.17***	0.25***	0.38***	N/A
	(0.042)	(0.040)	(0.092)	
	[1346]	[1339]	[1346]	

(a) Treatment variable for the 'Observed Design' and IV estimations: Owning a VIP latrine

Notes: The Original Design corresponds to the case where the treatment parameter is a dummy variable for being in the original treatment group, Phase A, regardless of when actual construction works were completed. The Observed Design corresponds to the case where the treatment parameter is a dummy for the villages where construction works ended before follow-up data collection. Standard errors clustered at the village level in parenthesis, except for the IV where the standard errors are bootstrapped. Sample sizes are in brackets. All the models include household fixed effects and the following covariates: the number of household members; number of household members under 5; number of elderly household members; sex of household head; age of household head; and dummy variables for the education level of the household head. A few outliers are dropped from the time collecting water regressions, in particular 11 observations are dropped from the time collecting water from all sources, and 9 from the time collecting water from main source. These observations were dropped because reported time collecting water exceeded 8 hours a day. We analyzed the results also if we dropped cases where time collecting water exceeded 3 hours a day and the results did not change substantially.

Source: Baseline and Follow-up IEMS Surveys

* p<0.05 ** p<0.01 *** p<0.001

Table 12. The effect of RWSSA on Intermediate Outcomes Correcting for Sample Attrition

Outcome	Original Design	Observed Design	Instrumental Variable	Matching
Any HH member had diarrhea (past 2 weeks)	-0.020	0.018	-0.030	-0.034
	(0.036)	(0.036)	(0.044)	(0.047)
	[1346]	[1346]	[1346]	[850]
Any HH member (5 or older) had diarrhea (past 2 weeks)	-0.010	0.029	-0.026	-0.026
	(0.029)	(0.026)	(0.039)	(0.034)
	[1346]	[1346]	[1346]	[850]
Any HH member (below 5) had diarrhea (past 2 weeks)	-0.084	-0.046	-0.044	-0.082
	(0.065)	(0.074)	(0.062)	(0.093)
	[581]	[581]	[581]	[360]
Any HH member (below 5) had >1 incidences of diarrhea (past 2 weeks)	-0.044	0.047	-0.025	-0.038
	(0.038)	(0.038)	(0.038)	(0.039)
	[577]	[577]	[577]	[357]
Any HH member (5 or older) had >1 incidences of diarrhea (past 2 weeks)	-0.030	-0.0099	-0.039	-0.035
	(0.018)	(0.016)	(0.024)	(0.018)
	[1346]	[1346]	[1346]	[850]
HH spent money on medical visit (incl. travel)	-0.012	-0.0076	-0.0062	-0.022
	(0.012)	(0.013)	(0.018)	(0.018)
	[1346]	[1346]	[1346]	[850]
Household member missed work in last two weeks for diarrhea	-0.0036	0.010	-0.0062	-0.021
	(0.021)	(0.019)	(0.014)	(0.018)
	[1346]	[1346]	[1346]	[850]

Notes: The Original Design corresponds to the case where the treatment parameter is a dummy variable for being in the original treatment group, Phase A, regardless of when actual construction works were completed. The Observed Design corresponds to the case where the treatment parameter is a dummy for the villages where construction works ended before follow-up data collection. Standard errors clustered at the village level in parenthesis, except for the IV where the standard errors are bootstrapped. Sample sizes are in brackets. All the models include household fixed effects and the following covariates: the number of household members; number of household members under 5; number of elderly household members; sex of household head; age of household head; and dummy variables for the education level of the household head. F-statistic for the 1st stage of the IV is 656.

Source: Baseline and Follow-up IEMS Surveys.

* $p<0.05$ ** $p<0.01$ *** $p<0.001$

Table 13. The effect of RWSSA on Long-term Outcomes Correcting for Sample Attrition

Outcome	Original Design	Observed Design	Instrumental Variable	Matching
Used time saved from water collection for work	-0.019	-0.0042	0.0026	-0.0030
	(0.014)	(0.014)	(0.0037)	(0.012)
	[1346]	[1346]	[1346]	[850]
Number of HH members who worked at least 1 hour in the past 2 weeks[(a)]	0.025	0.053	0.037	-0.044
	(0.090)	(0.093)	(0.13)	(0.12)
	[673]	[673]	[673]	[425]
Total Hours worked in the past 2 weeks for money (men)[(a)]	-6.65	4.01	-9.76	-10.8
	(8.42)	(8.56)	(12.3)	(11.2)
	[632]	[632]	[632]	[408]
Total Hours worked in the past 2 weeks for money (women)[(a)]	3.55	5.54	5.21	0.95
	(4.69)	(4.50)	(6.88)	(6.91)
	[636]	[636]	[636]	[403]
Total Hours worked in the past 2 weeks for money (per capita)[(a)]	-0.66	2.16	-0.96	-3.62
	(4.66)	(4.69)	(6.83)	(6.22)
	[606]	[606]	[606]	[391]
Total Hours worked in the past 2 weeks for money by all members 13+[(a)]	-2.19	10.1	-3.21	-7.83
	(11.2)	(11.2)	(16.4)	(15.0)
	[606]	[606]	[606]	[391]
Any man in HH worked for one hour or more in the past 2 weeks [(a)]	-0.045	-0.030	-0.065	-0.070
	(0.043)	(0.043)	(0.062)	(0.052)
	[632]	[632]	[632]	[408]
Any woman in HH worked for one hour or more in the past 2 weeks[(a)]	0.061	0.027	0.090	0.040
	(0.042)	(0.044)	(0.061)	(0.056)
	[636]	[636]	[636]	[403]
Any HH member older than 13 worked for one hour or more in the past 2 weeks[(a)]	0.0016	-0.027	0.0024	-0.026
	(0.053)	(0.053)	(0.077)	(0.064)
	[606]	[606]	[606]	[391]
HH experienced improved income in last month[(a)]	-0.0031	0.017	-0.0046	0.0057
	(0.024)	(0.023)	(0.035)	(0.029)
	[673]	[673]	[673]	[425]
Total cash income from all sources in 2012[(a)]	4225.3	7310.4	6195.9	6117.0
	(4360.1)	(5901.2)	(6393.6)	(6909.1)
	[583]	[583]	[583]	[363]

Notes: The Original Design corresponds to the case where the treatment parameter is a dummy variable for being in the original treatment group, Phase A, regardless of when actual construction works were completed. The Observed Design corresponds to the case where the treatment parameter is a dummy for the villages where construction works ended before follow-up data collection. Standard errors clustered at the village level in parenthesis, except for the IV where the standard errors are bootstrapped. Sample sizes are in brackets. All the models include household fixed effects and the following covariates: the number of household members; number of household members under 5; number of elderly household members; sex of household head; age of household head; and dummy variables for the education level of the household head.

[(a)] Data only available at follow-up. Does not include household fixed effects. Standard errors are clustered at the village level, except for the instrumental variable estimations.

Source: Baseline and Follow-up IEMS Surveys

* p<0.05 ** p<0.01 *** p<0.001

Bounds for treatment effects

In addition to the IPW results in this section we construct bounds for treatment effects following Karlan and Valdivia (2011). Note that these results correspond to bounds for the difference between the originally defined treatment and control groups at follow-up. Table 14-16 show lower and upper bounds under different assumptions for missing data. Columns 1 and 7 show the lower and upper bounds assuming the 'worst-case' scenario. In this case, for the lower (upper) bound, missing outcome data in the treatment group is imputed as the minimum (maximum) value of each variable in the observed treatment distribution, and missing outcome data for the control group is imputed as the maximum (minimum) value of each variable in the observed control distribution. The second scenario (columns 2 and 6) imputes missing data in the treatment group for the lower (upper) bound the mean minus (plus) 0.25 standard deviations of the observed treatment distribution, and missing data in the control group to the mean plus (minus) 0.25 standard deviations of the observed control distribution. The third scenario (columns 3 and 5) does the same exercise but with a 0.1 standard deviation. In column 4 the difference between (originally defined) treatment and control groups at follow-up are displayed for each outcome.

Table 14 shows results for short-term outcomes. Not surprisingly the bounds for the worst case scenario are quite wide, and we cannot discard there is no difference statistically significant in the outcome of interest between treatment and control for any outcome. However, when we look at the other bounds, significant differences can still be detected along the lines discussed in the main body fo the report.

Tables 15 and 16 show results for intermediate and long-term outcomes, respectively. Given that the unadjusted differences are not statistically difference from 0, it is not unexpected that the estimated bounds indicate that there is not a significant difference between treatment and control groups for any outcome of interest.[39]

[39] Note that this may seem inconsistent with our results for whether any woman worked at least one hour in the previous two weeks, as for this outcome we do discuss a positive and significant impact in the main body of the report. However, as we explain in subsection 6.2 we only find a positive impact for the instrumental variables specification. We are not familiar with a method that produces this type of bounds when using instrumental variables to control for selection into treatment, which is why we only produce bounds for the difference between the originally defined treatment and control groups.

Table 14. Bounds for Unadjusted treatment effects – Short-term outcomes

	Lower bounds			Unadjusted difference at follow-up	Upper bounds		
Outcome	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Short-term outcomes							
HH has improved water source	-0.0081	0.36	0.38	0.40	0.42	0.44	0.47
	(0.067)	(0.055)	(0.055)	(0.066)	(0.055)	(0.055)	(0.055)
Time spent collecting water per day (all sources)	-296.3	-49.0	-37.7	-30.1	-22.6	-11.3	155.1
	(16.1)	(5.59)	(5.57)	(9.89)	(5.63)	(5.74)	(13.4)
Time spent collecting water per day (main source)	-298.7	-50.4	-39.4	-32.0	-24.7	-13.7	154.3
	(16.4)	(5.65)	(5.64)	(9.99)	(5.71)	(5.82)	(13.9)
Percent of HH members using toilet	-0.0067	0.28	0.31	0.33	0.35	0.39	0.49
	(0.065)	(0.057)	(0.057)	(0.072)	(0.057)	(0.057)	(0.059)
Toilet used by all HH members	-0.20	0.14	0.16	0.18	0.20	0.23	0.28
	(0.057)	(0.043)	(0.043)	(0.052)	(0.043)	(0.043)	(0.044)

Notes: Eleven observations are dropped from the time collecting water from all sources, and 9 from the time collecting water from main source. These observations were dropped because reported time collecting water exceeded 8 hours a day.

Standard errors clustered at the village level in parentheses.

Source: Baseline and Follow-up IEMS Surveys

Table 15. Bounds for Unadjusted treatment effects – Intermediate outcomes

	Lower bounds			Unadjusted difference at follow-up	Upper bounds		
Outcome	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Any HH member had diarrhea (past 2 weeks)	-0.44	-0.056	-0.032	-0.017	-0.00099	0.023	0.042
	(0.045)	(0.022)	(0.021)	(0.031)	(0.022)	(0.022)	(0.023)
Any HH member (5 or older) had diarrhea (past 2 weeks)	-0.46	-0.051	-0.032	-0.019	-0.0064	0.013	0.017
	(0.046)	(0.017)	(0.017)	(0.026)	(0.017)	(0.018)	(0.018)
Any HH member (below 5) had diarrhea (past 2 weeks)	-0.75	-0.14	-0.071	-0.027	0.017	0.083	0.54
	(0.027)	(0.014)	(0.014)	(0.042)	(0.014)	(0.014)	(0.025)
Any HH member (below 5) had >1 incidences of diarrhea (past 2 weeks)	-0.77	-0.093	-0.045	-0.014	0.018	0.066	0.53
	(0.026)	(0.0095)	(0.0092)	(0.026)	(0.0090)	(0.0091)	(0.026)
Any HH member (5 or older) had >1 incidences of diarrhea (past 2 weeks)	-0.49	-0.050	-0.035	-0.026	-0.017	-0.0027	-0.0071
	(0.046)	(0.011)	(0.011)	(0.016)	(0.011)	(0.011)	(0.011)
HH spent money on medical visit (incl. travel)	-0.47	-0.022	-0.011	-0.0042	0.0030	0.014	0.0069
	(0.046)	(0.0088)	(0.0088)	(0.012)	(0.0089)	(0.0091)	(0.0089)
Household member missed work in last two weeks for diarrhea	-0.47	-0.020	-0.010	-0.0036	0.0031	0.013	0.0059
	(0.046)	(0.0073)	(0.0071)	(0.0095)	(0.0071)	(0.0073)	(0.0072)

Standard errors clustered at the village level in parentheses.

Source: Baseline and Follow-up IEMS Surveys

Table 16. Bounds for Unadjusted treatment effects – Long-term outcomes

	Lower bounds			Unadjusted difference at follow-up	Upper bounds		
Outcome	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Used time saved from water collection for work	0.0027	0.0027	0.0027	0.0027	0.0027	0.0027	0.0027
	(0.0026)	(0.0026)	(0.0026)	(0.0026)	(0.0026)	(0.0026)	(0.0026)
Number of HH members who worked at least 1 hr (past 2 weeks)	-1.96	-0.032	0.035	0.081	0.13	0.19	0.44
	(0.22)	(0.070)	(0.069)	(0.093)	(0.068)	(0.068)	(0.074)
Total Hours worked in the past 2 weeks for money (Men)	-300.7	-14.7	-8.11	-3.69	0.73	7.35	42.6
	(26.4)	(5.81)	(5.75)	(8.59)	(5.74)	(5.80)	(9.11)
Total Hours worked in the past 2 weeks for money (Women)	-128.0	-2.38	1.89	4.73	7.58	11.8	34.1
	(12.5)	(3.45)	(3.37)	(4.87)	(3.31)	(3.32)	(5.19)
Total Hours worked in the past 2 weeks for money (Per-Capita)	-198.6	-6.12	-2.37	0.13	2.63	6.37	48.8
	(16.6)	(2.82)	(2.78)	(4.28)	(2.76)	(2.78)	(7.54)
Total Hours worked in the past 2 weeks for money by all members 13+	-372.5	-12.9	-3.74	2.38	8.49	17.7	82.9
	(31.7)	(7.74)	(7.66)	(11.7)	(7.64)	(7.69)	(12.1)
Any man in HH worked for one hour or more in the past 2 weeks	-0.35	-0.084	-0.043	-0.015	0.013	0.054	0.22
	(0.042)	(0.032)	(0.032)	(0.044)	(0.032)	(0.033)	(0.039)
Any woman in HH worked for one hour or more in the past 2 weeks	-0.33	0.017	0.053	0.078	0.10	0.14	0.23
	(0.049)	(0.031)	(0.031)	(0.042)	(0.030)	(0.030)	(0.031)
Any HH member older than 13 worked for one hour or more in the past 2 weeks	-0.27	-0.037	0.011	0.042	0.074	0.12	0.36
	(0.042)	(0.036)	(0.036)	(0.053)	(0.036)	(0.037)	(0.043)
HH experienced improved income in last month	-0.43	-0.043	-0.020	-0.0056	0.0091	0.031	0.045
	(0.045)	(0.018)	(0.018)	(0.025)	(0.018)	(0.019)	(0.019)
Total cash income from all sources in 2012	-65500.2	-349.3	1975.2	3524.9	5074.6	7399.2	130945.7
	(6089.3)	(2990.2)	(2934.5)	(3487.5)	(2888.9)	(2877.1)	(24447.5)

Standard errors clustered at the village level in parentheses.

Source: Baseline and Follow-up IEMS Surveys

ANNEX F: BASELINE BALANCE TABLES

Table 17 below presents the results of t-tests to analyze balance between our outcome variables in the baseline sample. The regressions test the difference in means between households in the treatment and control villages at baseline, restricted to the panel households only. The results show a balanced sample at baseline for the panel households. The only statistically significant difference between treatment and control at baseline was the time spent collecting water per day, from the main source. In this case, the control households spent more time than the treatment, and this was significant at the 5% level.

Table 17. Baseline Balance Table

Indicator	Control		Treatment		p-value
	Mean	N	Mean	N	
Short Term Outcomes					
HH has improved water source	0.55	303	0.61	370	0.133
Percent of HH members using toilet	0.33	303	0.38	370	0.183
Toilet used by all HH members	0.23	303	0.25	370	0.726
Percent of households with VIP latrine	0.09	303	0.11	370	0.300
Time spent collecting water per day (all sources)	108.77	292	102.45	361	0.468
Time spent collecting water per day (main source)	90.06	291	75.31	362	0.047
Intermediate Outcomes					
Any HH member had diarrhea (past 2 weeks)	0.10	303	0.10	370	0.853
Any HH member (5 or older) had diarrhea (past 2 weeks)	0.08	303	0.07	370	0.761
Any HH member (below 5) had diarrhea (past 2 weeks)	0.05	116	0.08	156	0.314
Any HH member (below 5) had >1 incidences of diarrhea (past 2 weeks)	0.02	115	0.03	155	0.449
Any HH member (5 or older) had >1 incidences of diarrhea (past 2 weeks)	0.02	303	0.03	370	0.597
HH spent money on medical visit (incl. travel)	0.00	303	0.01	370	0.117
Household member missed work in last two weeks for diarrhea	0.03	303	0.04	370	0.737
Any member missed school/work more than once in past year	0.04	303	0.06	370	0.270
Any member missed school/work more than twice in past year	0.02	303	0.03	370	0.415
Long-Term Outcomes					
Used time saved from water collection for work	0.01	303	0.03	370	0.104

Source: Baseline IEMS

ANNEX G: FIRST-STAGE REGRESSION RESULTS FOR INSTRUMENTAL VARIABLE ESTIMATIONS

Table 18. First-Stage Regressions for IV Estimations

Variables	Household received treatment (TOT)	Household has VIP Latrine
Household Assigned to Treatment Village (ITT)	0.689***	0.482***
	(28.17)	(18.66)
Baseline/Follow-up Flag	0.0137	-0.0180
	(1.43)	(-0.76)
Number of household members	0.00144	-0.000657
	(0.22)	(-0.07)
Number of household members under 5	-0.0223	-0.0138
	(-1.33)	(-0.67)
Number of elderly household members	-0.0295	-0.0135
	(-0.80)	(-0.26)
Sex of household head (male)	0.0774*	0.0893
	(2.18)	(1.38)
Age of household head	0.00145	0.00144
	(0.30)	(0.22)
Age of household head (squared)	-0.00000743	-3.38e-08
	(-0.15)	(-0.00)
Household head has 0 to 4 years of education	0.0514	-0.0517
	(1.46)	(-1.41)
Household head has 5 or more years of education	0.0116	-0.0114
	(0.28)	(-0.22)
Constant	-0.120	0.0191
	(-0.84)	(0.10)
Observations	1346	1339

Standard errors are bootstrapped; t-statistics in parenthesis

* p<0.05 ** p<0.01 *** p<0.001

ANNEX H: ACCOUNTING FOR MISSING DATA FOR TOTAL TIME COLLECTING WATER

Both the baseline and follow-up surveys asked households to estimate the total amount of time they spent collecting water. This was an important short-term outcome for our analysis, as we would expect that having an improved water source closer to a household would decrease the amount of time households spent collecting water per day. We found that there were a much larger number of households with missing data for time collecting water at follow-up than at baseline. Out of our entire analysis sample, 326 were missing data for this indicator at follow-up, compared to only 33 at baseline. To investigate possible reasons why data may be missing at follow-up, we compared the skip patterns of the surveys between baseline and follow-up to see if there were any systematic reasons why a household may have skipped this question at follow-up but not at baseline. Though we found two differences in survey skip patterns between baseline and follow-up, when tabulating these questions at follow-up we found that the skip patterns would have accounted for less than 10 out of 326 missing values.

Though the reasons for the remaining discrepancies in missing values between samples could not be determined, we ran a probit model to see if having a missing value for time collecting water was statistically correlated with being in the treatment group. The results of the probit are given in Table 19. According to the model, being in the treatment group had a statistically significant correlation to having missing data for time collecting water.

Table 19. Probit Model to Investigate Missing Values for Time Collecting Water.

Variables	Time Spent Collecting Water is Missing
Treatment	0.189*
	(0.106)
Time spent collecting water per day (all sources)	0.000315
	(0.000469)
Number of household members	-0.0411*
	(0.0237)
Number of household members under 5	0.0273
	(0.0906)
Number of elderly household members	0.0403
	(0.149)
Sex of household head (male)	0.128
	(0.117)
Age of household head	0.0270
	(0.0245)
Age of household head (Squared)	-0.000228
	(0.000245)
Household head has 0 to 4 years of education	-1.280**
	(0.589)
Household head has 5 or more years of education	-1.289**
	(0.582)
Observations	653

Standard errors in parentheses; * p<0.10; ** p<0.05; *** p<0.01

Since there is a statistically significant relationship between being a treatment household and having missing data for this indicator, we generated inverse probability weights (IPW) to correct for any systematic differences between these households. The IPW weights households by the inverse of the probability that a household had missing data for water collection times. That way, households with higher probabilities of having missing data could be given larger weights in the regression. We then ran the regressions using these weights. The results of the weighted regressions are shown in Table 20 and reveal negligible changes in results for time spent collecting water compared to those in the main text using unweighted regressions.

Table 20. Weighted Regressions on Time Spent Collecting Water (all sources)

Outcome	Original Design	Observed Design	Instrumental Variable	Matching
Treatment	-31.58	-31.80	-48.04**	-21.27
	(17.00)	(16.78)	(15.29)	(20.78)
Observations	1098	1098	1098	698

Notes: The Original Design corresponds to the case where the treatment parameter is a dummy variable for being in the original treatment group, Phase A, regardless of when actual construction works were completed. The Observed Design corresponds to the case where the treatment parameter is a dummy for the villages where construction works ended before follow-up data collection. Standard errors clustered at the village level in parenthesis, except for the IV where the standard errors are bootstrapped. All the models include household fixed effects and the following covariates: the number of household members; number of household members under 5; number of elderly household members; sex of household head; age of household head; and dummy variables for the education level of the household head. Eleven outliers are dropped from the regressions. F-statistic for the 1st stage of the IV is 656.

Source: Baseline and Follow-up IEMS Surveys

* p<0.05 ** p<0.01 *** p<0.001

ANNEX I: ADDITIONAL SUMMARY STATISTICS

Table 21. Additional summary statistics

	Baseline		Follow-up			
			Control		Treatment	
	Mean	N	Mean	N	Mean	N
Type of Water Container Used		673		243		298
25 liter	0.7%		10.7%		8.1%	
20 liter container	83.4%		76.5%		78.9%	
10 liter	0.9%		10.3%		12.1%	
Other	14.7%		0.8%		0.3%	
Don't know	0.1%		0.4%		0.0%	
How many containers used	2.8	672	3.2	125	3.3	145
Who collects the water		661		258		304
Mother alone	39.5%		38.0%		40.8%	
Mother and daughter	23.0%		20.2%		23.4%	
Mother and son	4.5%		3.9%		2.6%	
Father	7.3%		4.3%		3.3%	
Other adult woman	7.4%		11.2%		10.5%	
Other adult man	4.7%		5.0%		6.6%	
Boys under 18	4.5%		5.8%		3.6%	
Girls under 18	8.5%		10.5%		8.2%	
Water vendor	0.6%		0.0%		0.3%	
Distance to water source (meters)	342	622	350	236	151	285
Respondent is satisfied with the toilet	41.0%	266	73.6%	106	22.5%	276

Source: Baseline and Follow-up IEMS Surveys

Table 22. Additional summary statistics – Hygiene habits

	Baseline		Follow-up			
			Control		Treatment	
	Mean	N	Mean	N	Mean	N
How frequently do you clean the toilet?		266		92		244
Daily or almost	28.6%		17.4%		37.7%	
Weekly or almost	44.4%		43.5%		41.4%	
Twice a month or almost	11.3%		5.4%		7.4%	
Once a month or almost	4.1%		5.4%		4.9%	
Less frequently	11.7%		27.2%		7.8%	
How frequently do you wash your hands after using the toilet?		673		298		367
Always	79.8%		72.1%		74.9%	
Sometimes	17.5%		20.5%		19.3%	
Never	2.7%		7.4%		5.7%	
How frequently do you wash your hands before eating?		667		298		367
Always	80.2%		66.4%		69.5%	
Sometimes	15.9%		22.8%		22.6%	
Never	3.9%		10.7%		7.9%	
How frequently do you wash your hands before eating food?		667		298		367
Always	79.9%		64.4%		68.7%	
Sometimes	15.6%		23.2%		22.1%	
Never	4.5%		12.4%		9.3%	
Has your household participated in hygiene promotion activities?		673		299		367
No	82.2%		75.9%		74.7%	
Yes some of the household	17.1%		21.4%		21.8%	
yes all the household	0.7%		2.7%		3.5%	

Source: Baseline and Follow-up IEMS Surveys

ANNEX J: CONSTRUCTED VARIABLES

Variable	Baseline construction	Follow-up construction
improved_water	Equal to 1 if B1_1 (main water source) is equal to 1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 15, or 18; 0 if otherwise	Equal to 1 if B1 (main water source) is equal to 1, 2, 3, 4, 5, 6, 7, 8, 9, or 12; 0 if otherwise
unimproved_water	Equal to 1 if B1_1 (main water source) is equal to 6, 14, 16, 17, 19, 20, or 21; 0 if otherwise	Equal to 1 if B1 (main water source) is equal to 10, 11, 13, 14, or 15; 0 if otherwise
toilet_improved	Equal to 1 if B38 (type of toilet) is equal to 1 or 2; 0 if otherwise	Equal to 1 if B46 (type of toilet) is equal to 1 or 2; 0 if otherwise
toilet_unimproved	Equal to 1 if B38 (type of toilet) is equal to 4, 5, or 6; 0 if otherwise	Equal to 1 if B46 (type of toilet) is equal to 4, 5, or 6; 0 if otherwise
toilet_none	Equal to 1 if B38 (type of toilet) is equal to 7; 0 if otherwise	Equal to 1 if B46 (type of toilet) is equal to 7; 0 if otherwise
toilet_all_use	Equal to 1 if B43 = 1; 0 if otherwise	Equal to 1 if B52 = 1; 0 if otherwise
toilet_not_used_elderly	Equal to 1 if B44 = 2; 0 if there is more than one over-65 person in household but B44 is not 2; missing if otherwise	Equal to 1 if B53 = 2; 0 if there is more than one over-65 person in household but B44 is not 2; missing if otherwise
toilet_not_used_children	Equal to 1 if B44 = 1; 0 if there is more than one under-5 person in household but B44 is not 1; missing if otherwise	Equal to 1 if B53 = 1; 0 if there is more than one under-5 person in household but B53 is not 1; missing if otherwise
toilet_percent	Divide B45 (number of people that use toilet) by household size; change to 1 if B43 (toilet is used by entire household) = 1; if value is still over 1 and B44=2, change to number of household members under 65 divided by total household size; if value is still over 1 and B44=1, change to number of household members over 5 divided by total household size; if B38 (type of toilet) = 7 (no toilet), change to 0; if value is still over 1, change to missing	Divide B54 (number of people that use toilet) by household size; change to 1 if B52 (toilet is used by entire household) = 1; if value is still over 1 and B53=2, change to number of household members under 65 divided by total household size; if value is still over 1 and B53=1, change to number of household members over 5 divided by total household size; if B46 (type of toilet) = 7 (no toilet), change to 0; if value is still over 1, change to missing
total_collect_time	For each of the three main water sources, multiply B29 (length of time, in minutes, it takes to get to water source) by B30 (number of times any HH member walks to get water per day), then add the three resulting values; change to missing for a couple of outliers (720 minutes in village 27, 600 minutes in village 48)	For each of the three main water sources, multiply B33_WS (length of time, in minutes, it takes to get to water source) by either B34_day_WS (number of times any HH member walks to get water per day) or B34_week_WS (number of times per week), dividing the latter by 7. Then add the three resulting values. Change to 0 if B25=0 (HH has water source in their home).
collect_main_source	Multiply B29_1 (length of time, in minutes, it takes to get to main water source) by B30_1 (number of times any HH member walks to get water per day)	Multiply B33_WS_1 (length of time, in minutes, it takes to get to main water source) by either B34_day_WS_1 (number of times any HH member walks to get water per day) or B34_week_WS_1 (number of times per week), dividing the latter by 7. Change to 0 if B25=0 (HH has water source in their home).
more_water_time	Equals 1 if B36 = 3 (HH spent more time collecting water compared to 12 months ago); equals 0 if B36 = 1 or 2; missing if otherwise	Equals 1 if B41 = 3 (HH spent more time collecting water compared to 12 months ago); equals 0 if B41 = 1 or 2; missing if otherwise
more_washing_time	N/A (necessary question not in baseline survey)	Equals 1 if B40 = 3 (HH spent more time washing compared to 12 months ago); equals 0 if B40 = 1 or 2; missing if otherwise
diarrhea_HH	Equals 1 if C1a_1 or C1a_2 equals 1 (a household member in either age category had diarrhea in past 2 weeks); otherwise, equals 0	Equals 1 if C1_ca_under5 or C1_cb_over5 (number of HH members of each age group that had diarrhea in past 2 weeks) is non-missing; otherwise, equals 0
diarrhea_HH_pct	N/A (necessary question not in baseline survey)	Divide sum of C1_ca_under5 and C1_cb_over5 by the total household size; change to 1 if value is over 1
diarrhea_under5_pct	N/A (necessary question not in baseline survey)	Divide C1_ca_under5 by number of under-5s in household; change to missing if HH has no under-5 members; change to 1 if value is over 1

Variable	Baseline construction	Follow-up construction
diarrhea_over5_pct	N/A (necessary question not in baseline survey)	Divide C1_cb_over5 by the number of over-5s in household; change to 1 if value is over 1
diarrhea_high_over5	Equals 1 if C17_1 is equal to 2, 3, 4, or 5 (over-5 HH member who had diarrhea in last 2 weeks had illness at least twice in past year); otherwise, equals 0	Equals 1 if C22_cx_over5 is equal to 2, 3, 4, or 5 (over-5 HH member who had diarrhea in last 2 weeks had illness at least twice in past year); otherwise, equals 0
diarrhea_high_under5	Equals 1 if C17_2 is equal to 2, 3, 4, or 5 (under-5 HH member who had diarrhea in last 2 weeks had illness at least twice in past year); otherwise, if HH has at least one under-5 member, equals 0; if not, change to missing	Equals 1 if C22_cx_under5 is equal to 2, 3, 4, or 5 (under-5 HH member who had diarrhea in last 2 weeks had illness at least twice in past year); otherwise, if HH has at least one under-5 member, equals 0; if not, change to missing
visit_cost	Sum of both C12 variables (cost of medical visit for under-5 and over-5 member with diarrhea)	Sum of both C16 variables (cost of medical visit for under-5 and over-5 member with diarrhea)
visit_cost_plus_travel	Sum of both C11 and C12 variables (cost of medical visit and travel for under-5 and over-5 member with diarrhea)	Sum of both C15 and C16 variables (cost of medical visit and travel for under-5 and over-5 member with diarrhea)
visit_cost_dummy	Equals 1 if visit_cost_plus_travel is non-missing; equals 0 otherwise	Equals 1 if visit_cost_plus_travel is non-missing; equals 0 otherwise
timeout_over5	Sum of C14_1 (days over-5 person with diarrhea missed work) and C16_1 (days that another HH members missed work to take care of over-5 person with diarrhea)	Sum of C18_ct_over5 (days over-5 person with diarrhea missed work) and C21_cw_over5 (days that another HH members missed work to take care of over-5 person with diarrhea)
timeout_under5	Sum of C14_2 (days under-5 person with diarrhea missed school) and C16_2 (days that another HH members missed work to take care of under-5 person with diarrhea)	Sum of C18_ct_under5 (days under-5 person with diarrhea missed school) and C21_cw_under5 (days that another HH members missed work to take care of under-5 person with diarrhea)
timeout_over5_dummy	Equals 1 if timeout_over5 is non-missing; equals 0 otherwise	Equals 1 if timeout_over5 is non-missing; equals 0 otherwise
timeout_under5_dummy	Equals 1 if timeout_under5 is non-missing; equals 0 otherwise	Equals 1 if timeout_under5 is non-missing; equals 0 otherwise
diarrhea_miss_12mo	Equals 1 if C18_1 or C18_2 (the sick HH member has missed school or work in the past 12 months) or C19_1 or C19_2 (another HH member missed school/work in past 12 months) is equal to 1, 2, 3, 4 or 5; equals 0 otherwise	Equals 1 if C23_cy_under5 or C23_cy_over5 (the sick HH member has missed school or work in the past 12 months) or C24_cz_over5 or C24_cz_under5 (another HH member missed school/work in past 12 months) is equal to 1, 2, 3, 4 or 5; equals 0 otherwise
time_saved_work	Equals 1 if B37_1, B37_2, or B37_3 = 1 (HH used time saved from collecting water to work); equals 0 otherwise	Equals 1 if B42_B_bcccc equals 1 or 2 (HH used time saved from washing/collecting water to work); equals 0 otherwise
hours_worked_men	N/A (necessary question not in baseline survey)	Sum of all A13_n values (hours worked in past two weeks for pay) when A4_d = 1 (member is male); value is missing if no one in household worked for pay
hours_worked_women	N/A (necessary question not in baseline survey)	Sum of all A13_n values (hours worked in past two weeks for pay) when A4_d = 2 (member is female); value is missing if no one in household worked for pay
number_workers	N/A (comparable question not in baseline survey)	Number of household members where A13_n is non-zero
hours_worked_nonschool	N/A (necessary question not in baseline survey)	Sum of all A13_n values where A6_f (age) is greater than 18
income_improved	N/A (necessary question not in baseline survey)	Equals 1 if A32_Ab (cash income last month) is greater than A33 (monthly cash income a year ago); equals 0 otherwise
income_2012	N/A (necessary question not in baseline survey)	Equals A35_Ad (total expected case income in 2012)
worked_men	N/A (no question on hours worked available at baseline)	Equal to 1 if A4 = "Male" and A13 reports 1 or more hours worked for any member of the household; 0 if otherwise
worked_women	N/A (no question on hours worked available at baseline)	Equal to 1 if A4 = "Female" and A13 reports 1 or more hours worked for any member of the household; 0 if otherwise
worked_nonschool	N/A (no question on hours worked available at baseline)	Equal to 1 if A6 shows the respondent is 13 or older, and A13 reports 1 or more hours worked for any member of the household; 0 if otherwise

ANNEX K: NORC'S RESPONSES TO REVIEWER COMMENTS TO THE MIDLINE IMPACT EVALUATION REPORT

Reviewer Name/ Institution	Page Number	Comment	Evaluator Responses
Steve Lowry	General	1. More explanation of who the water minders are, how they were selected, were they paid - if so, by who and how much?	The water minders were selected by the VHWC. Usually two are selected within each of the villages. They are encouraged to work side by side with the contractor to understand how the system is built and how it operates. The contractor is supposed to provide them with a toolkit to allow them to fix small repairs, such as leaks, taps, etc. They are not paid, however, the VWHC is expected to develop a plan to collect money monthly from households to pay for any future maintenance issues.
Steve Lowry	General	2. more explanation of the control group and what they received in terms of water or sanitation. Normally a control group has nothing done to it, but it seems in this program they had some facilities built.	Midline occurred in Nov-Dec 2012 with a second trip to the field in Apr 2013 to visit households that were missed in first attempt. As shown in Table 1 in the report, before Apr 2013 construction was completed in only one Control village, specifically in March 2013. Even in this case we consider that only one month is too little time for the program to have any impact. Along these lines, we can safely say that Control villages were not affected by any water construction before the midline.
Steve Lowry	General	3. Of the $164M MCC funds, how much was for this program?	$30MM. This has been included in the report.
Steve Lowry	General	4. Wasn't one of the long term outcomes to reduce poverty through economic development....?	Better health outcomes and more time available were supposed to affect labor and schooling outcomes. Long-term results are discussed now in the main body of the report (before, those results were in an Annex)
Steve Lowry	Page 15	5. Pg 15. Sounds like collecting data had some significant challenges. Author claims results of bungled data collection is negligible. Who is qualified to make this determination? When input data are questionable, then outcomes are also questionable. Not sure this is fully summarized with a clear conclusion - as in levels of confidence for the results.	Please specify which aspects of data collection seem particularly problematic so we can address them.

Reviewer Name/ Institution	Page Number	Comment	Evaluator Responses
Steve Lowry	Page 15 (cont)	In the revised report, footnote 24 presents a case for problems with the data.	This refers to the issue we found with the diarrhea incidence variable at midline. We agree this is concerning. This type of problem would have been detected in a DQR of the midline data but this work was delayed mostly because updating the evaluation design was prioritized. At this point we do not think would be productive to approach the BoS for clarification. That being said, in this particular case we believe that the solution we proposed is reasonable as for practical purposes the variable we propose using should capture the outcome we are interested in.
Steve Lowry	Page 16	6. pg. 1 6. Further confusion on keeping households straight. Indicates poor initial planning and training of people doing the questionnaires. With this amount of confusion, how valid are the results?	Please specify which aspects of data collection seem particularly problematic so we can address them.
Steve Lowry	Page 16 (cont)	Footnote 24 indicates that the responses were not clear and new questions had to be asked. And last paragraph on pg 16, which points out confusion on where treatment took place, or didn't take place.	Please see our response above on footnote 24. The discussion in this paragraph is simply about addressing construction delays in the impact evaluation, and the techniques that are of standard use to deal with these problems. However, both MCC and NORC acknowledges that there were issues, some of the preventable, with the contract structure, implementation planning, fieldwork, oversight, and delays in analysis due to additional design work that led to questionable data quality. As such, in the analysis, we have made statistical fixes to address these problems.
Steve Lowry	Page 17-19	7. Pg 17 - 19. Still more inconsistencies, etc are discussed, all trending towards negative impacts on the validity of the results. To the casual reader like myself, it appears that there were more problems than correct data collection. the author is spending a lot of time trying to find ways around the problem. Lesson learned - plan better in the beginning and train field staff better.	It would be perhaps more useful that the reader commented on which specific aspects of the evaluation methods are inconsistent. Instrumental variables and the type of 'Matching' we proposed are standard techniques in the empirical literature to address a problem as pervasive in program implementation as are construction delays.
Steve Lowry		8. Table 4. Is another reason for the drop in improved water source the lack of maintenance and failure of the system?	We cannot discard this hypothesis but why would this affect the control group more, and why was not observed before?
		For toilets, the text refers to "endline" whereas the table refers to "midline". Which is is?	Midline. This has been fixed.
		Was the study meant to look into the question of why toilet usage wasn't 100%, if the GoL funded toilets for all villagers? Same comment on collection of water....	Is this a question about the program coverage or the households' take-up?

Reviewer Name/ Institution	Page Number	Comment	Evaluator Responses
Steve Lowry	Page21	9. Table 5. Footnote on pg 21 indicates treatment resulted in more illness. Not the first time that more water has resulted in more illness - refer to USPHS studies on Navajo Reservation - which I recall showed the same as people were practicing poor hygiene and more water was contaminated and in contact with people - i.e. babies being bathed in contaminated water which was then not disposed of, or reused. Seems that when the results didn't coincide with what was expected the question was then changed to give "better" results. Again, seems like poor planning from the outset.	No. Please see the table and footnote. The change in how diarrhea was measured was conducted for both treatment and control group, so it cannot have the suggested effect on our estimates of the treatment effect.
Steve Lowry	Page21 (cont)	See pg 22. Village names could easily have been standardized if the problem had been recognized early. Pg 22. Seems that testing of water quality would have been a basic test so that it was clear the new source was not contaminated, or less contaminated than the old source. Also could have been more study (or reference to existing studies) on how water might get contaminated between the tap and the end user – as in using dirty cups to take water from a bucket, storing water in open buckets, etc.	We agree and this should probably be incorporated in evaluations of future interventions similar to this. It is worth considering though the budget implications such complex type of data collection would entail.
Steve Lowry	Page 22	10. pg 22. 2nd para notes no major impacts. One can ask if there were any positive impacts at all - and the footnote mentioned earlier would indicate a negative impact. The author should explain if the slight reductions shown between control and treatment are significant, or within the sampling error.	Which effects are statistically significant and which are not is discussed in the paper
Steve Lowry	Page 22 (cont)	See pg 21 highlight	As we say in the report, we do not find significant effects on diarrhea reduction at standard levels of confidence (5%)
Steve Lowry	General	11. Unfortunately, my conclusion is that the results from a poorly planned and poorly executed program are not of much value. What I get from the conclusion is that this program did little to improve health, though there was some reduction in time to collect water. Whether that is a function of the program, or other factors - such as more rain and more springs - is not addressed.	The program may have had no significant effect on diarrhea. The fact that incidence was relatively low at baseline makes it hard for any intervention to have a large effect, although it is worth saying that all parameters have the expected sign, except for a few of the parameters under the 'Observed Design' which suggests positive selection. We wouldn't underestimate the importance of the effect of the program reducing time collecting water, given the importance of this outcome and the size of the estimated effects. There is no reason to assume that this is not a consequence of the program given the methods implemented.

Reviewer Name/ Institution	Page Number	Comment	Evaluator Responses
Steve Lowry	General (cont)	Again see statement on pg 21.	We do not find significant effects on diarrhea reduction at standard levels of confidence (5%)
Lerato	Page 6	Adding a questionnaire is confusing because AMP is not a questionnaire.	Dropped
Lerato	Page 6	Consider deleting.	Done
Lerato	Page 12	There is a contradiction. If they were scheduled to be completed in September 2011 and they were completed in March 2011 they were fast tracked not delayed. Please check the dates.	We said construction *commenced* between December 2010 and March 2011
Lerato	Page 17	Two or six months? Check.	An additional column has been included so we think this is no longer confusing
Algerlynn Gill	Page 4	Please add an executive summary.	Done
Algerlynn Gill	Page 5	This sentence doesn't really flow from the preceding sentence—aside from the reference to time savings.	Changed
Algerlynn Gill	Page 5	Please use the official name: Rural Water Supply and Sanitation Activity, though you're welcome to abbreviate it, and please be consistent throughout	Done
Algerlynn Gill	Page 5	Was?	Ok
Algerlynn Gill	Page 6	Not completely accurate since the compact didn't invest in the training aspect directly. It should have been coordinated but that isn't the same thing.	Changed to 'MCC's WSP coordinated or invested' (instead of just 'invested')
Algerlynn Gill	Page 6 (cont)	I don't think MCA or Cowater actually coordinated the training components either. What evidence is NORC using to support this statement? One critique of the intervention might be that it wasn't better coordinated so I think we need to be clear on this point.	Our understanding is that CoWater was involved in PHAST training provision; in any case the point that is being made is that this was an intervention that encompassed not only water access but sanitation and hygiene training. We changed the reference to MCC and call it simply RWSSA, so the text reads now: "Often, program attention tends to focus on the delivery of water systems without simultaneous attention being paid to sanitation facilities and hygiene promotion; by contrast, the RWSSA included all three components, hopefully setting the stage for long-term impacts in reducing disease and improving the productive lives of Lesotho's citizens. "
Algerlynn Gill	Page 6	Source, e.g., Compact ITT or Compact M&E Plan.	Lesotho Table of Key Performance Indicators. (February 2013) - Added

Reviewer Name/ Institution	Page Number	Comment	Evaluator Responses
Algerlynn Gill	Page 6	Just reiterating that PHAST and Aftercare trainings were not funded by MCC as indicated in the footnote, though the rest of the footnote is accurate. If you're referencing the snacks and per diems, I suggest using different language.	Rephrased
Algerlynn Gill	Page 6	Also, CLO needs to be defined either in footnote or next paragraph.	Done
Algerlynn Gill	Page 7	Some happened so far in advance, I wonder if it was connected to the construction schedule at all. I wonder if the lack of consistency is at all related to the results we see.	Given the documented balance between treatment and control we doubt that differences in time exposure to training will make any difference.
Algerlynn Gill	Page 7	Cite	World Bank's Water Global Practice – Strategy http://www.worldbank.org/en/topic/water/overview#2 - Added
Algerlynn Gill	Page 7	Does NORC have the projected and actual costs for this activity? If not, I can try to track that down as useful context for the results we see. Also, can you put the results into context using the monitoring results and program logic? I can provide the ITT data for this. The idea would be that we achieved output targets, achieved reductions in time to collect water but aren't seeing changes in higher-order results.	The projected cost was $30MM. The report is already discussing the results in terms of short-term and intermediate outcomes. Now that we are integrating long-term outcomes in the discussion we think this should be clearer.
Algerlynn Gill	Page 7	Defined how?	This paragraphs was dropped.
Algerlynn Gill	Page 8	Define	Done
Algerlynn Gill	Page 8	I think the information in the table is interesting but I recall omitting it from the EDR out of concerns about possible re-identification. Is this info needed to replicate your analysis or necessary for any other reason? If not, we can check whether the DRB has any concerns about publishing before removing.	It is needed to replicate the Instrumental Variable (IV) and Matching results. In the IV the data is used to construct the variable that is instrumented and in the Matching the data is used to condition the regressions. Furthermore, not only the data in Table 1 is needed but the one in Annex C (which contains much more information). If this information cannot be published then we cannot incorporate the observed delays in the analysis.
Algerlynn Gill	Page 8 (cont)	We'll have you present your de-identification strategy to the DRB before we post the final version of the report online to make sure everyone is on the same page about actions needed to protect privacy.	Ok. I don't think the table summarizing the different processes can compromise anonymity but we will talk about it. The table in the annex was dropped in the latest version, let me know what you think.

Reviewer Name/ Institution	Page Number	Comment	Evaluator Responses
Algerlynn Gill	Page 9	Aftercare training is missing. Did you ever get access to that information from DRWS? If so, please indicate with a footnote or something what information you lack and what you've tried to do to obtain it. Without it, we don't really know if that part of the program was concluded.	We have included in this table the information that was already available in Annex C. Data reflects reality in 2013-2014.
Algerlynn Gill	Page 9 (cont)	I don't understand why this wouldn't reflect the reality as of Cowater's report. I do understand the information doesn't impact your analysis but it does help to paint a picture of whether the plans were ever completed and what the situation might be like if we were to return to the field.	The figures reflect the latest data sent by Cowater.
Algerlynn Gill	Page 9	Does the timing of these trainings raise any concerns? (three comments like this in this table)	It is possible that the impact of PHAST training depreciates over time. However, the fact that this was not randomized complicates evaluating this effect. That being said, the fact that treatment and control group were pretty balanced at baseline suggests that PHAST on its own may have very limited impact anyway.
Algerlynn Gill	Page 9	Referring to the spreadsheet from the supervisory engineer? If so, please clarify.	These spreadsheets were provided to us by Algerlynn Gill, who specified they came from Satish Menon.
Algerlynn Gill	Page 9 (cont)	Again, this was not an "MCC spreadsheet" but rather one prepared by MCC's supervisory engineer	The source is now: "MCC's supervisory engineer" Who we assume is Satish Menon
Algerlynn Gill	Page 9	Please note initial and revised targets. Let me know if you need that information.	If the initial plan was to provide VIP latrine for all households and treatment villages we would prefer to use this as the original 'Intent to treat', and for the actually observed prevalence of latrines the survey data. Along these lines, we do not think we need updated targets for the regression analysis.
Algerlynn Gill	Page 9 (cont)	I requested them as context about what was planned for the intervention, not for the analysis. Thanks for adding them to the revised report.	Ok
Algerlynn Gill	Page 10	How did NORC identify which indicators to measure?	NORC selected the indicators according to the causal models and hypotheses MCC was interested in testing (See NORC's Evaluation Mini-Report. January 2009)
Algerlynn Gill	Page 10 (cont)	Unfortunately, our hypotheses weren't very specific about the behaviors that should change, which is why I asked how NORC identified specific indicators—in other words, were the indicators selected based on the training that was done?	In terms of measuring change in behaviors, other than use of improved water source and toilet, the hypothesis of 'Greater hygiene awareness leads to improved hygiene behavior' was dropped anyway because it required the CTV method, so no indicator was constructed.

Reviewer Name/ Institution	Page Number	Comment	Evaluator Responses
Algerlynn Gill	Page 10	Please include a lit review, note what gaps if any the evaluation fills, note the evaluation type. For the lit review, please prepare a brief synthesis rather than a detailed presentation by study.	A revised version of the lit review include in previous reports has been included.
Algerlynn Gill	Page 10	This is only included for a few indicators; please flesh out	Done
Algerlynn Gill	Page 11	I would frame this differently. NORC proposed to use CTV. Unresolved issues related to exogeneity resulted in the EMC opting for the original evaluation methodology, which did not present the same concerns.	Footnote replaced with: Five of the 13 original hypotheses were supposed to be tested using Continuous Treatment Variable (CTV) approach. After discussing with MCC's Evaluation Management Committee (EMC) it was decided that it was preferable to drop the hypotheses that use this method and focus on the ones that could be evaluated using randomization as the key source of variation for identification of the treatment effect. As a result, no analysis using CTV is presented.
Algerlynn Gill	Page 11 (cont)	I think it's fair to say that NORC mapped hypotheses to the updated program logic diagram (some of which did overlap with the "original hypotheses") and **proposed** methods to test those hypotheses. These were presented to MCC for review and we did give the instruction referenced at left. However, saying these hypotheses were original or were "supposed to be tested using" CTV isn't quite accurate.	Footnote replaced with: Five of the 13 ~~original~~ hypotheses **in the Revised Evaluation Design Report** were supposed to be tested … (etc).
Algerlynn Gill	Page 13	To the extent all questions rely on the survey and progress reports, should the data sources not be the same throughout?	Yes, for simplicity we are dropping both columns on data sources and including a note at the end of the table.
Algerlynn Gill	Page 13	I wonder if we should tweak this and the reference in the next column. We really mean that they'll use water from a safe source since we have no way of knowing what happens with respect to how the water is collected and stored and whether it is indeed safe when used or consumed.	The underlying outcome we think is still that, but the indicator description was changed to: Degree to which household collects water from improved sources
Algerlynn Gill	Page 13	Drawing an intentional distinction between this and water system constructed indicator? If so, do we need to think about which is used in the subsequent regressions?	Yes. We discussed in which specific regressions we use availability of VIP latrine as the covariate of interest.
Algerlynn Gill	Page 13	Add an indicator for time spent washing clothes?	Response rate for this question is pretty low, which is why we can't use it.
Algerlynn Gill	Page 13 (cont)	Then why reference washing clothes here?	Agree, dropped.

Reviewer Name/ Institution	Page Number	Comment	Evaluator Responses
Algerlynn Gill	Page 13	Perhaps considering reframing the hypotheses above (#4,5) as "Program does…" or as "Access to X,Y,Z does…" since we aren't really testing for the relationships specified.	Done
Algerlynn Gill	Page 13	Specify that this is related to health (to distinguish from #3)	Done
Algerlynn Gill	Page 16	Omitting reference to health for simplicity?	Rephrased
Algerlynn Gill	Page 17	Did NORC validate that sampling was carried out according to plan?	During data collection, NORC's resident and local staff observed data collection at the start and during the survey. The sampling was being implemented according to plan. However, at some point in the fieldwork process, BoS opted not to visit some panel households, replacing them with new households at midline. However, since NORC wasn't privy to data extracts during the field period, we only detected these sampling issues till much later when MCA handed over the datasets to NORC. For this reason, BoS had to return to the field much later in April 2013 to administer the survey to all households interviewed at baseline in a given village. Other discrepancies persisted

Reviewer Name/ Institution	Page Number	Comment	Evaluator Responses
Algerlynn Gill	Page 17	Please flesh out discussion with required sample size; design omissions in sample; level of representativeness; strategy for absent respondents.	We added this at the end of the section: Given that the final panel sample is smaller than originally planned, it is important to discuss the potential consequences of this situation. Sample deterioration has two main implications. First, a smaller sample reduces the precision of the estimated impacts. This implies that we may find coefficients that are not significant, or only marginal significant, that with the original sample we would had found significant. The second problem is more serious because sample deterioration could be such that treatment and control groups are no longer comparable. Fortunately, as we discuss in more detail below, we do not find major differences between treatment and control groups in observable characteristics using the final sample panel, which shows that randomization was not compromised by sample deterioration. Finally, even if treatment and control groups in the final sample are balanced, sample deterioration could compromise the external validity of the results. It is worth saying that, in any case, this study was not going to produce results that were representative of a large population (like rural areas in Lesotho), because villages were selected for the study purposefully (as opposed to randomly), so the results are 'representative' only of the households in the selected villages. However, the panel sample (and the results derived from it) may not be representative even of the households in the selected villages due to sample deterioration. To address this possibility, in Annex J we use Inverse Probability Weights to correct for sample attrition. As we show, the results are not sensitive to this correction.
Algerlynn Gill	Page 17	Sanitation?	Done
Algerlynn Gill	Page 17	Are there any implications for the results?	The only case where we find evidence that this could constitute a problem is for time collecting water, we address this problem explicitly in the results section. A note was added.

Reviewer Name/ Institution	Page Number	Comment	Evaluator Responses
Algerlynn Gill	Page 17	When was this problem discovered? Could it have been identified beforehand?	Enumeration Area (EA) codes rather than villages were used to construct the household identifier variables, leading to duplicate ID codes for different households. Also, the village names recorded in the data could not always be linked back to villages in the sample, since villages often have multiple names. For similar reasons, there was also a discrepancy between the village names used in the intervention documents (e.g., the list of places where the rural water intervention was done) and the village names used in the sample (which simply came from the 2006 census). Essentially, village names are not unique or consistent enough to use to easily match places in different sources, or to match the sample against the collected data. On the other hand, EAs are problematic because individuals are not often aware of which EA they inhabit, and EA boundaries can be redrawn. Along these lines, improving the quality of the list of villages in census, and generate universal identifiers for all villages in the country, may be necessary to avoid this type of problems. The problem was identified during analysis phase; we did not discover the problem earlier, since, as explained above, NORC did not receive the data until well after the field period was completed.
Algerlynn Gill	Page 18	Does this mean HHs were interviewed more than once? Surveys were entered more than once? Or different interviews/HHs were given the same IDs?	The third one, different households had the same ID. A note was added.
Algerlynn Gill	Page 18	How?	After sorting by geographic location, the merge was conducting using household level data, like names. A note was added.
Algerlynn Gill	Page 18	As opposed to what other villages?	We clarified that in the IEMS data there are A and C villages, that take part in this study, and other villages that are surveyed for the other activities. We included this paragraph: A total of 871 panel cases were successfully merged, equivalent to 27 percent of the households surveyed at baseline. Note that this corresponds to all the households surveyed by the IEMS, which includes not only Phase-A and Phase-C villages, but also villages that take part in the studies of the health and water urban activities of the MCA-Lesotho Compact. When restricting the dataset to only households living in the Phase-A or Phase-C villages, there were 673 panel cases, equivalent to 71 percent of the A- or C-village households surveyed at baseline.

Reviewer Name/ Institution	Page Number	Comment	Evaluator Responses
Algerlynn Gill	Page 18	How does 871 HHs represent 27% and 673 HHs also represent 71% of the same baseline sample? Is there a distinction here I'm not following?	See previous answer
Algerlynn Gill	Page 18	Where does this situate us in terms of the power calculations and required sample size?	It is possible that sample reduction explains why we are not getting significant results (at standard levels of confidence) for diarrhea, given that all the indicators have expected sign. An analysis of power post-midline data collection could had been conducted but MCC instructed NORC not to conduct post-midline data collection analysis, while updates to and review of the evaluation design was being conducted.
Algerlynn Gill	Page 18 (cont)	MCC did provide the go-ahead to conduct the analysis, which is why you've produced the midline report. Thank you for adding the updated power calculations.	Ok About our previous response it is worth clarifying one more time that indeed there was no stop work order but rather a delay while the evaluation design was approved.
Algerlynn Gill	Page 18	Referring to the construction phases here? Maybe distinguish between village names used by DRWS and those used by BOS.	Done
Algerlynn Gill	Page 18	And no system had been devised to address this, right?	We understand the government has started working on the problem about a year ago, to standardize the names between DRWS and BOS villages.
Algerlynn Gill	Page 18	For the baseline HHs, do we know whether DRWS actually revisited them?	Revisited them for what purpose?
Algerlynn Gill	Page 18	For the midline, are these all the HHs that were visited erroneously?	It is possible that some of these households are part of the panel but it was not possible to merge due to the described identification issues.
Algerlynn Gill	Page 18	As a separate issue, I would still like to document what went wrong and how we could have avoided these challenges. As a matter of fact, some of the critical lessons learned from this evaluation are related to data collection. Let's discuss.	NORC did not have a contract with BoS directly. The BoS contract was with MCA; and, it was a time and materials contract. Payments were not linked to products and product quality. While NORC oversaw training and the start of data collection, and conducted observations during discrete points in the data collection, we were not present in the field during the whole field period and, hence, were not aware of sample alterations that occurred at various points in the data collection. Also, because NORC did not receive extracts of data during the field period, and did not receive the actual datasets till several months after end of data collection, we did not discover data quality issue till much later. To avoid this in the future, we would recommend not to do T&M contracts with data collection firms, and use of tablets if possible for data collection.

Reviewer Name/ Institution	Page Number	Comment	Evaluator Responses
Algerlynn Gill	Page 20	So the random assignment is not actually used when looking at results related to VIP latrines? What if HHs already had a VIP latrine? Can we say we are truly measuring the impact of the program related to VIP latrines? Has NORC run the analysis using the construction variable—is that what the original design model represents for the sanitation outcomes?	That is not correct, randomization is used in the context of estimating the impact of latrines when we use the IV approach. We use the random dummy to instrument having a latrine (model the probability of having a latrine as a function of random assignment), and then plug the predicted probability in the outcome equation. It is the same exercise that we do when using random assignment as an instrument for construction, but instead of dummy for having finished the works is a dummy for having a latrine. In the IV model we are using the random variation in having a latrine provided by treatment assignment as the covariate of interest, not simply the dummy for having a latrine, this is why cases where there was actually latrine before, or controls that get a latrine, are less of a concern. All the models called 'Original Design' use the randomization dummy. We use the raw construction variable in the models called 'Observed Design' and instrumented in the ones called 'Instrumental Variable' across all outcomes except the couple related to toilet usage, as in these cases we use having a latrine (for Observed design and IV).
Algerlynn Gill	Page 20	Why wouldn't the instrument for random assignment represent "the program," which includes both water and sanitation?	It does represent the program, and we present results for this way of looking at treatment provision. However, because the construction delays, if we stopped at the 'Original Design' we may be underestimating the impact of the program because we are basically saying that some hh were treated when in fact they were not. If the program has any effect, this would dilute the estimated effect. Using the IV is a way to correct this, because it incorporates the delays issue, but still exploits the randomization as the key identification strategy.
Algerlynn Gill	Page 21	Long-term only required 9 months, so why is NORC drawing this distinction here; the midline survey covered long-term results?	Long term outcomes were covered by the midline survey, the results are now in the main section of the report
Algerlynn Gill	Page 21	Can you also show some descriptive stats for other interesting variables, like B25/26, B27, B28, B46, B48, B50 (number in baseline instrument)?	Including an annex with these and other variables baseline descriptives
Algerlynn Gill	Page 21	Has NORC checked whether other data sources reflect such decreases in access?	We haven't researched other data sets
Algerlynn Gill	Page 22	Do we know if there were any epidemics during this time?	We haven't found any evidence of this.

Reviewer Name/ Institution	Page Number	Comment	Evaluator Responses
Algerlynn Gill	Page 22	How do these levels of water-related illness compare to what we'd expect? Would NORC conclude that this wasn't a major problem at baseline?	According to the DHS Lesotho, 2009, the Percentage of children under-five with diarrhea in the two weeks preceding the survey was 14% in Lesotho in 2009, so the figures we present are relatively close.
Algerlynn Gill	Page 22	Related to footnote 19 are all of these variables really referring to having sought treatment for diarrhea? If so, that needs to be clarified and a point made that this is a lower bound on estimates of illness.	Not exactly, as we explain in the footnote the dummy we constructed corresponded to whether there was an answer to the question on seeking treatment, not if they effectively sought treatment. So there is no clear reason for considering this a lower bound for incidence, as every respondent was supposed to answer if they reported having a member with this condition.
Algerlynn Gill	Page 23	This is labeled as "Matching" in the tables but there's no description of any matching procedures. Please clarify and include if relevant.	In the Methods section (p19) we included this clarification when we explain the fourth and final empirical method: "We call this method 'Matching', in the sense that we are restricting the sample to the districts where there are both treatment and control villages, and dropped the districts were, effectively, there are only untreated villages, and no treated villages to 'match' to.".
Algerlynn Gill	Page 24	What does NORC make of this—one significant result for each of these variables, across different specifications?	Our preferred specification is the IV because it exploits the randomized treatment assignment but acknowledges the delays that occurred. We highlight this by modifying a little our discussion on the effects on time spent collecting water that know reads: "[…] Similar results are observed for time collecting water from the main source. However, for this outcome, the estimate for our preferred specification, the IV model, is only marginally significant, indicating that the effect of the program on time savings from collecting water from the Main source is smaller than the effect associated with All sources of water." Also, we could argue that gains in wellbeing derived from spending less time collecting water should be considered comprehensively rather than depending on the type of water.
Algerlynn Gill	Page 24	Define when introducing the different specifications earlier (or in column headings)	The references to TOT and IV-TOT were incorrect and replaced by 'Observed Design' and 'Instrumental Variable', which are discussed in the methods section and presented in the tables.

Reviewer Name/ Institution	Page Number	Comment	Evaluator Responses
Algerlynn Gill	Page 26	Policy implications? What have we learned with these results?	We added at the end of the conclusions: In terms of policy implications, the results described in this report imply that this type of program can have major impacts on households wellbeing via reductions on time spent collecting water, but limited effects on outcomes that may seem more important, like diarrhea incidence. Furthermore, even if household members spend less time collecting water, it is not clear that this will translate in greater labor force participation, as labor outcomes may depend on more factors than just greater labor supply. On the other hand, it would be a mistake to undervalue the importance of reducing time collecting water. It is possible that these time savings will have effects on outcomes that cannot be observed by an instrument like the one fielded in the context of this evaluation. For example, more available time for children could have an effect on time studying, which could have an effect on test scores. More time studying and greater academic achievement will presumably translate in greater opportunities for children in the future.
Algerlynn Gill	Page 26	Another limitation was lack of plan to test water quality. Related to my comment above, we can't speak to whether people consumed clean water because we aren't testing quality at point of source and point of consumption.	We added this in the conclusion as an additional explanation for why no effects on diarrhea were found: "The modest impact on diarrhea may also be because the quality of consumed water may have not improved significantly. Because water was not tested at the point of consumption, we cannot document the extent to which water quality actually improved or not."
Algerlynn Gill	Page 26	I wasn't sure what to make of this result. Does this statement require more nuance?	In what sense?
Algerlynn Gill	Page 26 (cont)	This was related to the fact that time savings were only found when looking at all sources, rather than the main source and weren't significant across the various specifications but you've made the case now that all sources should count more than the main source and that the IV model is what NORC is focused on.	Ok

Reviewer Name/ Institution	Page Number	Comment	Evaluator Responses
Algerlynn Gill	Page 26	Please include next steps or recommendations for future analysis. This is where NORC can describe any publications you might pursue or mention your recommendation of what could be learned with another round of data collection.	With respect to more data collection, we don't think it would be very useful to run another wave of data collection at this point. If we start planning an endline data collection now we will probably be going to the field in 2017, more than 5 years after the intervention, perhaps too much time after both treatment and control have been treated to pick up any effects. An alternative is to use other data to analyze this problem, like the Demography and Health Survey (conducted in 2014), but sample size may be a problem if there is not too much overlap between the study villages and the villages surveyed in DHS; another alternative is the census that is being collected this year (supposedly). Because is an RCT we don't need at baseline strictly speaking, so maybe that's something we can look at.
Algerlynn Gill	Page 26	Please also include references.	Done
Algerlynn Gill	Page 27	The survey includes a number of other really interesting variables in addition to those I listed earlier. Here are some others: functionality of water source; reliability of supply; whether people used time savings for productive purposes; source of toilet; B47-50, 52-53; why people don't seek treatment, C3. It could be interesting to use some of this data to flesh out the need at baseline as well as our understanding of what changed since.	We could present summary statistics and briefly discuss the figures, is that what you are asking for?
Algerlynn Gill	Page 27 (cont)	As mentioned at left, it could be interesting to use those variables to flesh out the baseline picture more and look at what has changed. However, you don't need to do this.	I agree. It would be interesting to use this data in order to explore mechanisms underlying the final results, especially if we want to think of questions for an endline evaluation.
	Page 36	Control?	Done

Reviewer Name/ Institution	Page Number	Comment	Evaluator Responses
	Page 36	Does this not conflict with NORC's initial plan for a 9-month lag? How much time do you think is necessary and why would this differ significantly from the time required for health outcomes? What about the relationship to short-term and intermediate outcomes? If we aren't seeing significant changes in amount of time spent collecting water, being sick, caring for the sick, then the channels through which we'd expect these variables to change just aren't there. I suggest moving this analysis into the main body of the report and framing results using the program logic. Where is it breaking down? What other variables can we use to get at these things, e.g., did distance to water source change even though we didn't consistently find time savings in collection times?	We moved the long-term results to the main body of the report. And framed the results using the program logic, although our presentation differs from what you are suggesting, in particular at the end of the results section we added: Given that the program has no significant effects on health outcomes, perhaps it is not surprising that no significant long-term effects can be estimated. However, long-term outcomes were supposed to be affected not only via health improvements but also greater time availability. Given that the (short-term) results presented before showed evidence that the program has reduced the amount of time household members spent collecting water, perhaps it is puzzling that no effects are found on labor outcomes. It is possible that time availability does not translate into better labor outcomes because the latter are not restricted by time availability but by other conditions, like the labor market itself. This could be because the local labor market cannot absorb much more labor supply, especially when we consider that time savings maybe were observed by most people in each treatment village, rather than just the surveyed households. This type of general equilibrium effects should be addressed in future research. It is worth highlighting that while the effects on labor outcomes may be negligible, some household members, specifically children and teenagers, could be using available time in different learning activities, which could have an effect on academic achievement. In sum the RWSSA has had substantial short-term effects. The program is associated with greater access to improved water sources and greater toilet use, and less time spent collecting water. In terms of intermediate outcomes, no significant effects are found for diarrhea incidence, although the signs for most of the analyzed variables (and in the case of the IV, all of them) indicate a negative correlation between the program and diarrhea incidence and its costs. Finally, no effects were found for long-term outcomes, namely labor outcomes. It is possible that, while the program freed up time that could have been used to get more work, labor market conditions prevented this from occurring. In any event, the value of having more time available should not be disregarded, even if it does not reflect directly on labor outcomes.
	Page 39	I suggest switching order for consistency with earlier tables.	Done

Reviewer Name/ Institution	Page Number	Comment	Evaluator Responses
Algerlynn Gill	Page 32	Protected?	'Well, neighbor' is classified as an improved water source. Note that only two households in the sample list this option as one of the available sources of water.
Algerlynn Gill	Page 32	Unimproved in all cases?	To be conservative 'Other' is classified as unimproved. Only 24 households reported this as a source of water.
Algerlynn Gill	Page 32	Is a VIP not an improved toilet?	This variable considers piped sewer system and septic tanks. VIP latrine is coded as a separate variable. Note that this variable, the VIP latrine one, is the one we use in our analysis.
Algerlynn Gill	Page 32 (cont)	I the VIP variable just not in the constructed variables annex?	The definition is included now.
Algerlynn Gill	Page 35	Can you include the variable for having a VIP at baseline?	Done
Sello Sefali	Page 5	The investment covered construction of 250 water supply systems and up to 30,000 VIP latrines, benefitting a population of approximately 150,000. The Compact funding covered construction of 90 water supply systems and 9,807 VIP latrines while the GoL funding covered construction of 160 water supply systems and 19,287 VIP latrines.	Program output updated using cited references.
Sello Sefali	Page 9	O&M manual?	There is reference to this manual in the Cowater activity completion report. Is this not correct?
Sello Sefali	Page 9	Is this the case? Needs verification.	Clarified that the program was assigned at the water system level rather than at the village level (and the water system can serve more than one village).
Sello Sefali	Page 19	There are 250 systems and each system has 5 villages on average.	Clarified that the program was assigned at the water system level rather than at the village level (and the water system can serve more than one village).
Sello Sefali	Page 34	What could be the reasons. Is it diarrhea?	Yes the analyzed intermediate outcomes are all related to diarrhea incidence and costs. As we discussed in the report it is possible that the estimated effects of the program on diarrhea incidence are not statistically significant because diarrhea had a relatively low incidence rate at baseline to start with. The modest impact on diarrhea may also be because the quality of consumed water may have not improved significantly. Because water was not tested at the point of consumption, we cannot document the extent to which water quality actually improved or not.

Reviewer Name/ Institution	Page Number	Comment	Evaluator Responses
MCC Evaluation Lead	General	Please describe in more detail what the installed "water systems" consisted of in the introduction. It is important for the reader to be able to understand the theory of change.	Included the underlined text in the Program Description subsection: During the course of the project 269 water systems were implemented, 19 more than the 250 originally targeted. Water system modalities included boreholes with hand pumps, solar powered pumping systems, gravity-fed spring catchment systems, and electric pumping systems. Each system encompassed between 2 and 5 villages. In treated villages standpipes were placed according to the village's demand and how far apart houses were from each other. According to the audit by the Project Management and Construction Supervision, for the most part households in treated villages had a standpipe within 150m of distance, which is the DRWS standard for service.
MCC Evaluation Lead	General	When stating the outputs of the project (water systems and VIP latrines), please differentiate between what was completed with compact funds by CED and what was completed with GoL funds. This can be in a footnote, but we don't want the compact monitoring data to contradict with this report.	This was added to the ES and the project description: RWSSA originally included 250 rural water supply points and 10,000 VIP latrines and had a budget of $30.2 million (18 percent of the $164-million Water Project in the Compact). In order to increase the coverage of VIP latrines in participating villages, MCC subsequently increased the budget to $40.1 million and the Government of Lesotho (GOL) contributed $17.1 million to RWSSA. In addition, the target for VIP latrines coverage was increased from 10,000 to 27,245 in the Lesotho M&E Plan. When the Lesotho Compact ended in September 2013, 175 water systems (70% of the target) and 29,352 VIP latrines (108% of the target) had been installed. Implementation continued post-Compact with approximately $5.3 million of additional funding from the GOL; ultimately, 250 water systems (100% of the target), and 31,768 VIP latrines (117% of the revised target), were completed. The total cost of RWSSA, including MCC and GOL funding during the Compact and after, was approximately $60 million.
MCC Evaluation Lead	24	This sentence has a typo: For these 34 villages – which are part of Phase Arev– the time of exposure to treatment before controls began receiving treatment between in January 2013 range from 10 to 19 months.	Edited as follows: For these 34 villages – which are part of Phase Arev– the time of exposure to treatment before controls began receiving treatment between in January 2013 ranges from 10 to 19 months.
MCC Evaluation Lead	36	Typo in this sentence: This is somewhat puzzling as more available time (due to reductions in time collecting water) could had translated into more time working.	Edited as follows: This is somewhat puzzling as more available time (due to reductions in time collecting water) could have translated into more time working.

Reviewer Name/ Institution	Page Number	Comment	Evaluator Responses
MCC Evaluation Lead	36	Is 43 minutes time savings in line with the literature / program expectations? Is that amount of time savings really likely to result in increased labor participation? I'm just wondering if the notion of time saved in collecting water translating to economic productivity increases is realistic.	Probably not, but the fact that it is not obvious what the result would be is what makes this an important empirical question. As it was shown, time savings seem to have had impacts on the extensive margin (more women working at least one hour) but not on the intensive margin (more hours worked).
GSI	30	This is the first place where the results has been segregated by gender. Considering women often bear the responsibility for managing water supply and sanitation in most households in Africa, it is expected that more results on the impact of the RWSSA project on women will be shown in this report..	Use of time data at the individual level was only collected for working for money, which are the results we are presenting in the report. While collecting use of time data for each individual would had been interesting, it would had also been very time consuming, requiring possibly to collect diary data at the individual level.
GSI	35	It would have been good to have more of the findings segregated by gender.	See previous response
GSI	36	It is really difficult to see from these findings if there are gender-based differences. For examples, is there a greater time savings by women than men? Any information on what the women did prior to the new water schemes and whether they increased their original activities even if it is not paid labor? There are so many unanswered questions about the impact the RWSSA have had on women who are usually responsible for water collection/management and sanitation maintenance.	See previous response. Also, a footnote saying that, according to the data, females are in charge of collecting water in almost 80 percent of the households, has been moved to the main body of the report (p. 37).

Reviewer Name/ Institution	Page Number	Comment	Evaluator Responses
M&E Lead	General	Since MCC has decided not to pursue a third round of data collection, can we replace the references to "midline" with "follow-up" instead? Feel free to include a footnote or something that explains that three rounds of data collection were initially envisioned but the plan changed. Here's the language we're using in the Summary of Findings, though this will be tweaked once we reach agreement on the terminology, that you're welcome to crib: "The evaluation design contemplated an endline round of data collection to explore the trajectory of results over time. However, having reviewed the midline results, NORC and MCC agreed that an additional round of data collection was unlikely to improve our understanding of the program impacts measured at midline. NORC proposed two alternative endline studies for MCC's consideration: (1) assessing physical and cognitive outcomes for children who were age 0-2 years at baseline or (2) assessing the current status of supported infrastructure, hygiene-related behaviors, and community support structures. Both options represent interesting opportunities. However, since the first option goes beyond the outcomes initially targeted, coupled with the weak health effects found at midline, MCC did not consider it a promising investment. The second option is quite relevant but given that MCC and MCA-Lesotho had very little involvement in the complementary training on behavior change and sustaining the rural water infrastructure, the second option does not represent a direct evaluation of the MCC-funded intervention which makes it less relevant as a standalone study. This evaluation is complete and there are no next steps."	We replaced midline with follow-up. In a few cases where we talked about 'midline evaluation results' (like in the title of the report) we just say 'evaluation results'. We added the following as a footnote in p 22: "The original evaluation design contemplated a third round of data collection to explore the trajectory of results over time. However, having reviewed the follow-up results, NORC and MCC agreed that an additional round of data collection was unlikely to improve our understanding of the program impacts measured at follow-up."

Reviewer Name/ Institution	Page Number	Comment	Evaluator Responses
Raymond Guiteras	General	Weaknesses. It would help to be more transparent about how precise the null results (no effect on X) are. That is, are the estimates precise enough that we can say with some confidence that the intervention did not affect X appreciably, or just that the estimates are sufficiently noisy that even a plausible effect size could avoid detection? Recommended changes. Precision as noted above.	The evaluation did not find significant impacts for intermediate and (most) long-term outcomes. For intermediate outcomes we doubt that the reason was the lack of power. As we show in Annex D the resulting design was not underpowered to identify an impact on diarrhea (any hh member). In the case of diarrhea-related outcomes what we think was the driving factor was that diarrhea prevalence was low to begin with. For labor outcomes for women, on the other hand, it is possible that we failed to find impacts on hours because of reduced power. In p. 37 we are adding this text: "Another possible explanation is that for this outcome we are underpowered as a consequence of the sample deterioration discussed in previous sections of this report. As we show in Annex D, the impact that the program needed to have on hours worked for women for this sample to estimate treatment effects with acceptable precision was higher than the MDES originally planned. Therefore, we cannot discard the possibility that, with a bigger sample, we would have been able to estimate treatment effects with acceptable precision. In any case, given the impact on time savings, any impact the program may have had on hours worked for females would probably have been relatively small."
Raymond Guiteras	General	Recommended changes. The manuscript is admirably transparent about the attrition problem. However, the checks and fixes noted (e.g., IPW), while correct, only address whether attrition was a function of observables. It would be helpful to provide some rough estimates of how much bias from attrition related to unobservables may have affected the results. See Karlan and Valdivia, REStat, 2011, for some bounding methods.	We produced bounds following Karlan and Valdivia and included this analysis in the same annex where we present our IPW correction. In the main body of the report right after we discuss the IPW results (section 4.3, p25) we added: "...we follow Karlan and Valdivia (2011) and construct different sets of bounds for the treatment impacts, in order to assess the extent to which sample attrition may be biasing the results; we find that for most of the outcome the estimated bounds do not change the conclusions derived from our main specifications."

Reviewer Name/ Institution	Page Number	Comment	Evaluator Responses
Raymond Guiteras	General	Recommended changes. It would be an excellent service if the authors included a "Lessons Learned" section detailing what they would have done differently in retrospect. In particular, I would like to know why it was difficult to anticipate the village-matching problem, how they might have detected it at the time, and what they would do differently to avoid it. Along the same lines, it would be great for the authors to discuss whether their estimates were as precise as they had expected and if not, what they might do differently at the design stage, power calculations, etc.	We have included this at the end of the data proccessing section: "The main circumstance that made it difficult to preserve the panel of households over time was that the names of the village, the unit at which treatment status was assigned, were not useful as unique identifiers: there are many common names used for different villages, sometimes in the same district, and some villages have multiple names that do not resemble each other. In retrospect it would have helped to assign every sampled village a permanent and unique ID code to be reused for each round, and integrate GPS from the beginning to make sure interviewers go to the right village regardless of its name. Some of these limitations could have been tackled during the follow-up fieldwork; however, because NORC did not receive extracts of data during the field period, and only received the actual datasets several months after the end of data collection, most of these issues were discovered much later."

Reviewer Name/ Institution	Page Number	Comment	Evaluator Responses
Raymond Guiteras	General	Strengths. I am very strongly in favor of the process evaluation the authors have proposed (AMP). Weaknesses. In principle, I like the idea of looking at effects on young children. However, I need to be convinced that the resulting estimates will be precise enough to be informative. There are a few reasons I am skeptical: first, the large attrition noted in this interim report; second, my understanding is that these outcomes are inherently noisy, measured with error and require highly trained surveyors to measure; third, the "first stage" (effect on sanitation and water) was significant but not enormous, which would limit the plausible effect size on these outcomes. Recommended changes. AMP: I would like more specifics on the qualitative research, and hope that specialist qualitative researchers will be involved (from, for example, anthropology). Detailed technical documentation for the proposed option 2a, addressing my comments above. As a related note, I am somewhat unclear on what is being proposed in Option 2b. To maintain randomization, you would need to sample from the population of households residing in the community at baseline – otherwise, differential selection bias becomes an issue again. Also, you'll need to check whether differential fertility or child mortality could be large enough to lead to bias.	We addressed this and other concerns in a document where we describe in more detail these two options. That being said, given MCC has decided not to move forward with any of these alternatives, for the sake of brevity, we refrained from elaborating more in this report about this topic.