IMPACT EVALUATION DESIGN REPORT

MCA TANZANIA – WATER SECTOR PROJECT

Elaborated by:
Mike Duthie[1], Jeffrey Alwang[2], and Charles Pendley[3]

Submitted to the
Millennium Challenge Corporation

Social Impact, Inc.
June 15, 2012

[1] Social Impact
[2] Virginia Tech
[3] Independent Consultant

TABLE OF CONTENTS

ACRONYM LIST

CI	Confidence Interval
CMD	Coliform Microbial Density
DAWASA	Dar es Salaam Water and Sewage Authority
DAWASCO	Dar es Salaam Water and Sewage Corporation
DD	Difference-in-Differences
DHS	Demographic and Health Survey
DMA	District Metering Area
DMO	District Medical Officer
ERR	Economic Rate of Return
FCR	Free Chlorine Residual
FGD	Focus Group Discussion
GIS	Geographic Information System
GIZ	Detsche Gessellschaft für Internationale Zusammenarbeit
GPSM	Generalized Propensity Score Matching
IE	Impact Evaluation
IV	Instrumental Variable
KII	Key Informant Interview
MCA	Millennium Challenge Account
MCA-T	Millennium Challenge Account/Tanzania
MCC	Millennium Challenge Corporation
mg/L	Milligrams per Liter
MLD	Million Liters per Day
MORUWASA	Morogoro Urban Water Supply and Sewerage Authority
NBS	National Bureau of Statistics
NPS	National Panel Survey
NRW	Non-Revenue Water
NTU	Nephelometric Turbidity Units
SI	Social Impact
SOW	Scope of Work
TWSP	Tanzania Water Sector Project
UNICEF	United Nations Children’s Fund
UPS	Uninterruptable Power Supply
WHO	World Health Organization

INTRODUCTION

Water Activity Objective

Water Sector Project

The $65.2 million water sector project comprises three specific interventions: (i) expanding the capacity of the Lower Ruvu water treatment plant serving Dar es Salaam; (ii) increasing production and improving water quality in Morogoro; and (iii) providing support to improve management and increase revenue of the water authorities in Dar es Salaam and Morogoro[4]. The objective of the water project is to increase investment in human and physical capital and reduce the prevalence of water-related diseases.

Activities

Initially the Water Project encompassed the following components:

1. Lower Ruvu Plant Expansion: expanding the capacity of the Lower Ruvu water treatment plant serving the Dar es Salaam area, from about 180 million liters per day (MLD) to approximately 270 MLD;
2. Morogoro Water Supply: improving water supply in Morogoro through rehabilitating a water treatment plan and improving water transfer in the existing distribution network. The overall interventions will increase the production from the current 18 million liters per day to 33 million liters per day; and
3. Non-Revenue Water: improving system efficiencies in Dar es Salaam through reduction of non-revenue water via reduction in physical leaks and commercial losses.

Subsequently, due to cost and budget constraints, the non-revenue water (NRW) component in Dar es Salaam was scaled back to include a series of studies and mapping, training relevant DAWASCO staff and procurement of relevant equipment. No NRW activities were planned for Morogoro.

Beneficiaries

According to the MCA-T Monitoring and Evaluation Plan (April 2012), beneficiaries are estimated as the sum of existing and new customers by 2027. Customers include residential, industrial and commercial connections. Using the 2027 projections and assumed annual population growth rates, SI estimated the population benefitting from the Water Project by city during the 15-year period from 2013-2027, as shown in the following Table 1:

Table 1: Estimated Number of Water Project beneficiaries by City (2013-2027)

City	Year		
	2013	2020	2027
Dar es Salaam	1,261,077	1,805,829	2,585,898
Morogoro	90,816	140,025	215,961
Total	**1,351,893**	**1,945,854**	**2,801,859**

An average annual population growth rate of 5 percent is assumed for Dar es Salaam and 6 percent for Morogoro during the period from 2013 to 2027. This assumption is consistent with

[4] Although the majority of the third component was cancelled.

recent population growth rates in Dar es Salaam and Morogoro. It is recommended that the above population projections be adjusted when the results of the upcoming 2012 national population census are available, most likely by mid-2013.

Objectives and Project Logic[5]

The main objectives of the water sector projects are (i) to increase investment in human and physical capital and (ii) to reduce the prevalence of water-related disease in order to reduce poverty through economic growth. In order to effectively evaluate project impact and achievement of these objectives, the evaluation design must consider outcomes and short and medium-term objective results. It must also identify how these outcomes and results are distributed among focal population groups (by gender and by household economic status) and the mechanisms by which these outcomes and results objectives could be reached. Indicators of outcomes and short- and medium-term objectives are presented in Table 2. Measurement of these indicators is discussed in more detail below.

Table 2: Expected Results

Result	Expected Impact	Gender Specific Impact	Differential Impact on Poor?
		Outcomes	
Water service coverage	Positive	N/A—household-level effect	Bigger impact on poor
Water service quality	Positive	N/A—household-level effect	Unclear
Water quality	Positive	N/A—household-level effect	Unclear
Water consumption	Positive	N/A—household-level effect	Bigger impact on poor
Water expenditures	Negative	N/A—household-level effect	Unclear
Water security	Positive	N/A—household-level effect	Bigger impact on poor
		Short-term Objectives	
Water-borne disease related morbidity	Negative	Unclear	Bigger impact on poor
Human capital accumulation	Positive	Bigger impact on women	Bigger impact on poor
		Medium-term Objectives	
Mortality	Negative	Unclear	Bigger impact on poor
Economic activity	Positive	N/A—household-level effect	Bigger impact on poor
		Compact Goal	
Poverty	Negative	N/A—household-level effect	Bigger impact on poor

Note: Indicators for these results are included in Table 3

The immediate outcome of the expansion of the water supply in Dar es Salaam, conditional on completion of the new transmission main (as discussed below), and Morogoro is an increase in availability of water for households across the utilities' coverage areas. As increased volumes of water flow through piped water system, water access should increase. The project logic posits increased access to occur through both increased continuity of service and also through increased number of customers or connections. However, since the project does not support the extension of the distribution network, increases in customer numbers must be developed through one of three pathways:

[5] Detailed MCA-T project logic chart presented in Annex I

1. Increased network investment by DAWASA as they have additional water available to sell and see potential returns to infrastructure investments
2. Households currently in the network but unconnected because of cost or perceived lack of water availability
3. Inactive customers who previously did not have water through the connection

The works in both sites are also expected to have an effect on the quality of water accessible to households in the distribution networks. The Morogoro works are expected to have direct effects on the quality of water in the distribution network, while in Dar es Salaam, households may have increased access to better quality water as they substitute utility provided water for unimproved sources. Access to cleaner water is expected to result in decreased water-borne diseases, including diarrhea.

As a result of increased water access, the amount of time spent fetching water decreases, and household members can allocate this time to more productive activities, including income-generating activities. Moreover, with increased availability of water across the city, the cost of water should decrease, anticipated to lead to increases in water consumption. Households will likely depend less on water vendors[6], leading to decreased prices, as vendors face increased competition, and reduced water expenditures (with potential negative side-effects on livelihoods for water vendors). Another critical potential area of impact is in water security or vulnerability. Water security is related to water service quality and is defined by water volume, water quality, and consistency of access. A water secure household should have enough clean water year-round to ensure its members survival, health, and productivity (UNICEF 2002). Through increasing water availability and access and decreasing water costs, the program is expected to increase water security, which is measured both through households reporting of water shocks and through lower defensive expenditures due to this increased water security.

The potential health outcomes of increased access to clean, piped water, particularly the lower prevalence of water-related disease, also permit an increase in the time beneficiaries engage in productive activities. Directly, adults can increase participation in income-generating activities as water-related sickness occurs less frequently. Indirectly, adults will spend less time in opportunity costs, such as caring for sick children. It is important to note that while potential impacts on health-related outcomes is part of the project logic, and will be measured, it is highly unlikely we will be able to detect statistically significant changes in the incidence of water-related disease given the very large sample size requirements, which are discussed further in later sections of this report. The expected result of both the expanded water supply and increased water quality is the reallocation of time to activities that increase household income. This assumes that increased availability of water from the utility will improve the quality of water consumed, a critical hypothesis. Households without connections typically rely on water vendors or community connections, most of which are connected to the utility distribution network. However, storage and transport procedures may reduce water quality, or alternatively, for

[6] The GIZ Urban Water Poverty study estimates that 68% of the urban poor in Tanzania rely on informal water service providers who charge on average 13 times the tariff of utilities. According to NPS data, about 0.9 (3.5) percent of Morogoro households rely on water vendors as their primary source of drinking water in the rainy (dry season); 9.2 (12.6) percent of Dar es Salaam households rely on vendors as their primary source of drinking water. Unfortunately, the survey does not have detailed information on secondary and other sources of drinking water.

households that rely on new connections, increase regular interruptions in service which may reduce water quality through the leakage of effluent into empty pipes.

This impact evaluation will also examine the impact of project activities on investments in human and physical capital. With regard to human capital, the expected increase in water supply and decrease in water-related disease both increase the likelihood that children will attend school, as it overcomes previous obstacles such as household chores (collecting water) or sickness. Though less certain, it is also possible that adults will perceive increased returns to sending children to school if they are healthier. These factors will be measured using a household survey and a treatment effects approach will be used to establish a causal link between increased access to water and human capital investment. For adults, businesses may find it less risky to invest in employees (through expanded hiring or additional training) when they require fewer sick days or less time off to care for sick children. Moreover, with additional disposable income (through reduced water expenditures and time off work) as well as increased returns to investment (again from improved health), households and businesses are expected to increase investments in productive assets. Information on changes in these investments will be gained using qualitative interviews as well as an asset register in household surveys. The potential links between water sector improvements and the longer term objectives and goal, including increased human and physical capital investment, reduced mortality, and increased income) are less obvious than the short-term objectives (e.g. households with access to more water, lower incidence of water-related disease, etc.), are observed over a much longer time horizon, and are expected to represent much smaller effect sizes. All of these make establishing attribution difficult. Nevertheless, the proposed evaluation design will aim to provide a thorough and nuanced understanding of project impacts, but, the focus will be on measuring impact on the short-term objectives.

A direct relationship is expected between the number of households and businesses using improved water sources, increased per capita water consumption, reduced prevalence of water-related diseases and increased investment in human and physical capital. For these relationships to hold, a number of antecedent and intervening variables, which are seen as necessary conditions, must be considered. These are shown in greater detail in Table 3:

Table 3: Antecedents and Intervening Variables

Logical Parameter	Antecedents	Intervening Variables
Increased number of house-holds and businesses using improved water sources	Increased availability of treated water	Accessibility Reliability and adequacy of supply
Increased per capita water consumption	Increased availability of treated water	Accessibility Cost of water and ability of vulnerable populations to pay Reliability and adequacy of supply
Reduced prevalence of water-related diseases	Increased availability of treated water	Hygienic use of water No contamination along distribution channel Sanitary conditions Disposal of wastewater Promotional/awareness- raising activities
Increased investment in human and physical capital	Human capital investments depend on access to schooling activities, reductions in time spent hauling water, and reduced incidence of water-related diseases. Physical capital investments depend on business climate, access to capital, interest rates Public investment Entrepreneurial skills	Reduced time hauling water, reduced prevalence of water-related diseases Promotional activities Rates of return

Critical Assumptions

The above causal chain relies on increased supply of treated water to beneficiaries. However, this is contingent upon two critical assumptions:

1. Water Availability: Works in Morogoro and Dar es Salaam are designed to increase the potential output of treated water from the utility. However, if, for example during a drought or during the dry season, water intake to the treatment plant is constrained – as droughts may increase pressure on the pipe system as alternative sources of water become unavailable – we would not expect to see an increase in output of treated water. Based on discussions with MCA-T and the utilities, we do not expect this to be a significant issue.
2. Completion of the New Transmission Main in Dar es Salaam: The existing transmission main in Dar es Salaam is insufficient to transmit the additional output of the new water treatment plant. Therefore, DAWASA will not utilize the additional capacity until the plant can be connected to a new transmission main. The status of the development of the new transmission main is unclear, although it appears that the new transmission main may not be completed until well after the completion of the treatment train and possibly not until very close to or after the Compact closeout.

Both of these critical assumptions will be monitored throughout the evaluation. While the first assumption is continuous (i.e. even small amounts of additional water available would allow for increased output) and therefore less likely to jeopardize the project logic, the second assumption is discrete, and until it is connected, we would not expect to measure any change in outcomes in Dar es Salaam. Accordingly, the IE component in Dar es Salaam is contingent upon clear plans and a timeline for completion of the transmission main. This second assumption does not affect Morogoro.

Literature Review[7]

As noted above, the primary goal of the Tanzania Water Sector Project (TWSP) is to improve human capital and economic production and to decrease the prevalence of water-related disease. In support of these goals, a wide literature holds that improved access to clean water in the developing world is associated with significant health gains, particularly among children (Jalan and Ravallion 2003, Merchant et al. 2003, Fink et al. 2011). More broadly, the literature generally concludes that interventions increasing access to improved water have positively impacted their target populations, particularly by reducing occurrence of diarrhea and other water-related illness (Waddington et al. 2009). Evidence also indicates that increased water access can have positive economic benefits, particularly by increasing household savings and freeing up funds and time for other pursuits (Galiani et al. 2008). The proposed impact evaluation for TWSP will explore whether these general findings hold true in Tanzania and will also tease out the key causal mechanisms that drive the anticipated changes.

Salient research highlights a number of confounding variables that must be considered as we proceed with this analysis. To begin with, the MCC investment in Tanzania is designed to expand water production, so benefits to the population will depend upon expansion in water availability. Transmission mains, connection rates, and distribution mechanisms are all major intervening variables that will factor into our analysis of impact on the population. In addition, literature suggests that household income and education levels are highly correlated with health outcomes and must be controlled for (Jalan & Ravallion 2003, Lee et al. 1997, Novak 2011). Maternal education levels have been strongly linked to child health, demonstrating the importance of considering this factor in any health analysis (Desai & Alva 1998). Seasonal variation is another important factor upon which health and economic outcomes may vary substantially. Significant differences exist in time spent collecting water, employment, and expenditures between the rainy and dry seasons in Tanzania. Diarrhea incidence is significantly lower during the dry season, as well, making measurement significantly more difficult at that time. These important seasonal variations must be considered in any comprehensive study of the Tanzania water sector.

The literature on the impacts of improved water supply highlights two broad competing methods for conducting the type of evaluation proposed. The preponderance of similar studies conceptualizes treatment as a binary variable. In other words, researchers categorize water sources into a binary variable, either “improved” or “unimproved.” For example, in a rural system, treatment households have access to water from a borehole while comparison households are left as they were before. These newly improved households are easily categorized as treatment units, while the remainder can be classified as a control group (see Jalan and Ravallion 2003).

But these binary studies are plagued with two critical limitations. First, as Novak explains in her evaluation of a Senegalese water program (2011), binary models do not account for other water sources that may still be used. Novak demonstrates that piped water into a dwelling does not significantly reduce the incidence of child diarrhea as compared to other water sources,

[7] The literature review presented in the body of the paper synthesizes relevant academic research. But SI obtained more than 150 documents and completed a thorough document review. The results of the document analysis are outlined in Annex III

highlighting the importance of accounting for all water inputs when considering health outcomes. Binary classifications fail to account for these additional water inputs, reducing their effectiveness. Second, access to water is often continuous. In other words, even if a household has pipes or other newly installed water access nearby, constraints on the water system's capacity may influence actual water availability in different places. One household may have substantially less water access then a comparable household simply because of system capacity constraints. Water quality can also vary over a continuum, and water quality can vary by source and also changes as water moves from its source, to storage, to ultimate consumption. These realities of water distribution may make improved-unimproved binary classification inappropriate for many research situations.

These limitations to binary studies of water impact are especially salient the case of the proposed TWSP IE. All of the study's subjects already had access to some level of water – the question is whether *more* water and *better* water will cause measurable improvements. We fundamentally assume that increased water production will reach the beneficiary population through a functioning transmission main with which households can connect, either directly or indirectly, relatively easily. With this in mind, nearly all households can access water, but the system's capacity constraints will likely give different households access to different quantities of water. Therefore, we will be looking at both changes in the portfolio of water consumed by a household (assuming a shift towards more improved sources) as well as possible increases in total water consumption. With these considerations in mind, we outline a plan to invoke continuous or generalized propensity-score matching and instrumental variable regression analysis to more accurately measure the effects of water access on Tanzanian households.

Notably absent in the literature on water impact are impact evaluations of urban-focused water projects. Most evaluations have been rural-based, partially due to the relative ease of performing a rural water-section impact evaluation when compared to its urban counterpart. Rural projects typically involve providing water to populations who previously had negligible access to clean water, while urban water reforms involve many more clients and widely varying levels of water access, creating complicated measurement challenges. The TWSP will add substantial value to the literature here. In addition, studies on water-related illness in Tanzania are few, particularly since epidemics like HIV/AIDS have commanded the bulk of international research attention for the past two decades. However, in 2010, Napacho and Manyele studied drinking water quality in Temeke District. They discovered that the chemical parameters of water sources did not meet WHO and Tanzania Bureau of Standards (TBS) standards, suggesting that current water quality levels will need to be carefully established during the baseline study. The TWSP analysis will likely shed light here that can be built upon by future researchers.

EVALUATION QUESTIONS

Evaluation Type

> **Water Availability, Access, and Consumption**
>
> Throughout this report and study, we will distinguish between water availability, access, and consumption:
> - *Availability* refers to water supply, irrespective of access rights;
> - *Access* refers to the ability to obtain water; and
> - *Consumption* refers to the individual decision to use water.

After reviewing available data and discussing options with MCA-T and other stakeholders, SI recommends that that an impact evaluation (IE) be implemented on the Tanzania Water Sector Project. The project was designed with a clear cause-and-effect relationship in mind: improved water infrastructure should increase water use and improve health. Evaluation questions were developed in collaboration with MCA-T, MCC, sector stakeholders and the SI evaluation team. These questions clearly center on impact, requiring the evaluators to measure changes in water access, record important health indicators, evaluate economic impacts, and so forth. A less-intensive program evaluation would insufficiently address the appointed questions, and a rigorous IE methodology is necessary to adequately explore the effects of the TWSP on various aspects of Tanzanian life.

Evaluation Questions

After consultation with the MCA-T, MCC, stakeholders, and SI, the proposed evaluation questions are:

1. What is the project's impact on water supply at the utility level?
    a. Does this create additional customers?
2. What is the project's impact on the availability of, and access to, water, especially in terms of:
    a. Reduced time to collect water
    b. Changes in domestic and/or commercial sources of water
    c. Vulnerable groups' access to water
3. What is the project's impact on consumption patterns of water at the household level?
4. What is the project's impact on water quality, both at the source, along the distribution channel, and ultimately at the point of consumption?
5. What is the project's impact on health, particularly on the incidence of diarrhea for children under five?
6. What is the project's impact on poverty and income?
    a. In particular, what is the project's impact on household expenditures relative to doctor visits, illness, time use, etc.?
    b. Do households decrease water expenditures as a result of the project?
7. Do households increase investment in physical and human capital as a result of increased access to water?
8. Do the project's benefits and costs accrue differently to men and women (and other important sub-groups)? If so, how? What is the reason for these differences?
9. What effect does the program have at levels above the household, including on businesses, schools, and health centers?
10. What are the unintended (positive or negative) results of the project?
11. What is the likelihood that results will be sustained over time?

12. What is the cost effectiveness or re-estimated economic rate of return (ERR) based on realized benefits and costs of the project?
13. Process-related questions:
    a. Was the MCC investment implemented according to plan?
    b. What challenges were encountered? How were the challenges addressed?
    c. What are the lessons learned from the design and implementation?
    d. What variations in this activity might be worth considering in the future?

Given the wide range of evaluation questions, different evaluation and data collection methods must be utilized to respond to the full set of questions. Importantly, most questions will require multiple methods to be fully addressed. Questions 1-8 will principally rely on the primary quantitative IE design, but will also require additional, contextual information generated through qualitative methods, as described in greater detail below. Questions 8-10 will be primarily addressed through qualitative methods, including detailed interviews with a sub-sample of beneficiaries and non-beneficiaries. These qualitative methods will also be used to complement the quantitative impact analysis to deepen our understanding about why we observe what we observe. Questions 11-12 will rely on analysis of project records, and in the case of Question 11, comparing those with data and analysis generated through the IE. The process related questions, while not a focus of the IE, are critical for being able to interpret results.

By answering these questions credibly, the evaluation will not only provide a strong basis for understanding the effects of the TWSP, but will also fill significant gaps in the evaluation and project design literature regarding effectiveness of urban water programs, providing strong evidence for programmers and policy makers in the sector. Moreover, by demonstrating the use of relatively newly-developed (in the case of generalized propensity scores) quasi-experimental or econometric techniques to estimate the impact of continuous treatments, this study represents a significant contribution to the evaluation literature.

METHODOLOGY

General Approach

The impact evaluation will use a combination of evaluation designs and data collection methods. First, recognizing differences between timing of project completion and data availability in Dar es Salaam and Morogoro, we propose to employ slightly different approaches (mainly distinguished by differences in data) for each site. Second, recognizing the inherent complications in conducting an assessment with as many intervening and confounding variables as are present in our case, we propose triangulating results within each project with different estimation techniques. Our broad approach is to exploit variability in intensity of treatment; the water investments will affect, either directly or indirectly, virtually all residents in Dar es Salaam and Morogoro, but households will benefit differentially depending on their starting conditions (availability of water) and their position along the distribution grid. As a result, a continuous treatment approach is necessary.

We propose to triangulate using generalized propensity score matching (GPSM) combined with difference in differences (DD) (as it allows controlling for the change in non-observable variables), and instrumental variables (to control for potential sources of selection bias) to measure the change in individual- and household-level outcome variables within the "area of influence" of the projects. This analysis will be conducted on a panel survey of households. A second proposed form of triangulation is to collect qualitative data from different sources, including through key informant interviews (KIIs) and focus group discussions (FGDs), as detailed below. These interviews will help capture impacts on businesses, water vendors and public facilities—institutions that will not be covered by the household surveys. The qualitative information will deepen our understanding of the nature and sources of change and further help us interpret the quantitative analysis. The various methods of impact analysis considered are detailed in Table 4:

Table 4: Types of Impact Analysis for MCA-T Water Investments

Broad Methodology	Level of Analysis	Focal Outcomes & Data Sources	Data source	Evaluation Design	Information Provided	Potential Limitations
Quantitative	System-wide	Changes in health, mortality	Health reporting system	Before-After	In aggregate, what changes are observed?	Difficult to attribute changes to intervention
Quantitative	System-wide	Changes in economic activity (business start-ups)	NBS	Before-After	In aggregate, what changes are observed?	Difficult to attribute changes to intervention
Quantitative	Health District	Changes in reported diarrheal diseases	DMOs	Before-After	Disaggregated changes in health outcomes	Difficult to attribute changes to intervention
Quantitative	Health District	Changes in reported diarrheal diseases	DMOs & DMA data from DAWASCO	DD based on high versus low impacts over space	Disaggregated changes controlling for time-invariant area-specific factors	Some factors will change; high-versus low impacts is a rough approximation
Quantitative	Household	Changes in household-level indicators	Household survey	DD combined with GPSM	Differences controlling for time-invariant household-specific characteristics; estimate dose-response relationship	Depending on measurement of the treatment, there may be endogeneity problems
Quantitative	Household	Changes in household-level indicators	Household survey	Instrumental variable estimation of water uptake; estimation of dose-response relationship with endogeneity of water uptake controlled for	Differences controlling for time-invariant household-specific characteristics; estimate dose-response relationship; overcomes endogeneity problem	Heterogeneous effects may complicate estimates; rigid assumptions about “structure” of decisions
Qualitative	Business	Changes in business level outcomes, including revenue	KII, FGD	Before - After	In aggregate, what changes are observed?	Difficult to attribute changes to intervention
Qualitative	Individual	Gender outcomes	FGD, Household survey	Before - After	In aggregate, what changes are observed?	Difficult to attribute changes to intervention
Qualitative	School	Changes in school attendance	KII, Document review	Before - After	In aggregate, what changes are observed?	Difficult to attribute changes to intervention
Qualitative	Health Centers	Changes in number and severity of water-related illnesses seen	KII, Document review	Before - After	In aggregate, what changes are observed?	Difficult to attribute changes to intervention
Qualitative	Utility Managers	Changes in water usage, perceived health outcomes, perceived business uptake	KII, FGD	Before-After	In aggregate, what changes are observed?	Difficult to attribute changes to intervention

Preliminary Considerations

Continuous Treatment

During the Social Impact (SI) scoping mission in March 2012, the team spent considerable effort clarifying the nature of the change expected from the MCA-T investments. In both cities, the investments will lead to increases in water supply, and these increases will be felt throughout the utility (DAWASCO and MORUWASA) service areas. In both utilities, staff was unable to identify areas within their service areas where impacts were expected to be greatest. They agreed that impacts would be felt throughout the city, but were reluctant to identify even low-, medium- or high-impact areas. The reason for this generalized impact is that the distribution systems are interconnected. Due to this interconnectedness, for example, increased water flowing through the Lower Ruvu treatment plant would substitute for water currently being supplied from the Upper Ruvu in Dar es Salaam. Thus, although the supply change will occur in a single part of the system (Dar es Salaam has three main sources of water supply), impacts will be felt everywhere. The Morogoro improvements are occurring in two treatment plants, but impacts are likewise expected to be diffuse.

A second source of impact will be through the change in quality of water. In Morogoro, enhanced water treatment is part of the MCA-T investment, and this will directly affect the quality of water entering the system. While the Dar investments do not include improved treatment, delivered quality is likely to change indirectly (in Morogoro as well). Less rationing and fewer periods with no pressure will reduce infiltrations of contaminated water through cracks in pipes and improve the quality of water delivered through the system. This quality change is expected to vary depending on the location. Indeed, in areas where people relied on alternative sources for water (for example, boreholes or tankers taking directly from the utility), introduction of an irregular supply, which may be contaminated during outages, may actually yield decreases in water quality.

Under these circumstances, the "treatments" are conceptualized to be the change in availability of water and change in quality of water due to the MCA-T capacity investments. The nature of the treatment causes two challenges for the impact assessment: (i) methodology for creating a counterfactual; and (ii) measurement of the treatment.

<u>Methodological challenge</u>: The continuous nature of the treatment dictates the appropriateness of certain methods. Given the circumstances, in the face of potential biases, it is impossible to identify a specific control group: virtually everyone is both cities is affected in one way or another. This, combined with the obvious differences between Dar es Salaam, Morogoro, and other cities, means that pure controls cannot be found; that is, we could not identify alternative cities which might serve as an experimental controls. Therefore, in order to estimate a counterfactual the evaluation will exploit differences in the continuum of treatment using a continuous treatment approach.

<u>Measurement challenge</u>: Since the treatment is not binary and we will rely on differences in levels of treatment, it is important to measure it as precisely as possible. The behavioral hypothesis relating the treatment to changes in outcomes is clear: increased availability of water stimulates changes in behavior (such as consumption) resulting in outcomes (such as time

savings, income and health effects). All else equal, availability at any one point, depends on the amount flowing into the system, how the water is routed by opening and closing valves, and off-take (including sales by the utility and losses of non-revenue water). Water access is then determined by availability and resultant household behavior: determinants of household water consumption. Unfortunately, these determinants include both observable (income, household size, etc.) and unobservable components. These unobservable components are also likely to be correlated with unobservable determinants of outcomes of interest. While we may be able to model and control for part of this consumption decision, there is likely to be some residual unobservable component. If we compare outcomes of households with differing levels of consumption, we introduce a selection bias to our measurement of the treatment effect that will be difficult to fully identify or control for. Therefore, measuring availability of water is a critical component of measuring the treatment, which is central to being able to attribute varying outcomes with different levels of treatment. Consumption is not sufficient.

As a part of the NRW study, DAWASCO established 23 fairly well-defined District Metering Areas (DMAs)[8], and the metering instrumentation for the DMAs should be modified to measure both flow and pressure, giving a relatively accurate measure of availability at each metering point. GIS-based or engineering modeling could then be used to estimate supply for neighborhoods or households within each DMA; these estimates would be based on readings at the metering point, distance from the metering point (possibly adjusted for known points of off-take), and pipe characteristics between the metering point and the household/neighborhood in question[9]. The model will be used to measure the exogenous component of the treatment—the change in availability of water. The advantage of such a measurement scheme is that the treatment, as measured this way, is outside the household's control reducing possible sources of selectivity bias. It is also closest to the concept in question—the treatment is a change in the quantity and quality of water available. The validity of this approach relies on access to a good-quality GIS data for the water distribution network and on the integrity of the DMAs. We are confident, but not certain[10], this data exists for Dar, but it is unlikely available for Morogoro. It also requires the use of bulk metering equipment, including equipment purchased by the MCA-Tanzania. Bulk metering in equipment is available in Dar es Salaam, while in Morogoro the installation of bulk metering equipment is planned but not complete. Hence, a different approach is required for Morogoro.

Three alternatives to meter-based measurement of treatment level are: (i) measurement of water availability at the household; (ii) measurement of uptake (water consumption) at the household, and (iii) measurement of uptake averaged over a neighborhood or a community.

The first option could be constructed using questions similar to those in the NPS. For those who have connection to a piped water supply, the NPS asks how many days per week, on average,

[8] District Metering Areas are defined water distribution areas where the amount of water entering the distribution area is measured. The inflow is compared with billing data from the same sub-area to get an estimate of water losses. Consultation with Jacobs Engineering indicated that the DMAs in Dar are relatively tight, with minimum amounts of water entering the sub-network without being recorded. Achieving "tighter" DMAs would be prohibitively expensive.

[9] An alternative would be to establish metering points within neighborhood enumeration areas. At a sufficiently small geographical area, these neighborhoods could represent tight DMAs. This option is, however, cost prohibitive.

[10] The team has seen maps made from such GIS data but has been unable to get the underlying data. See the trip report for more detail on these challenges.

they receive water, and how many hours per day. If responses to these questions are reliable, they would provide a good measure of exogenous water supply. However, we expect a certain level of bias, as households may be unlikely to know how often water is available. This measure could be averaged over local networks to get measures of availability for piped connections and others in the area. We propose that these household-level responses be supplemented with community- or area-level information provided by water utility area agents. Each DAWASCO service area office has commercial staff who are responsible for billing in service blocks. These agents have extensive knowledge of water supply conditions within their service blocks and could accompany survey enumerators to provide information on days and hours of water service within the area. The main advantage to this service area approach is that it will measure availability exogenous to household decisions. Relying on household respondents may produce responses of questionable reliability, but if we join the household survey responses with expert opinions from the area agents, reliability will increase.

The second option (household estimates of consumption) has the advantage of being relatively easily measured with standard surveying practices. Its main disadvantages are that it is endogenous to household decisions, i.e. consumption could be plausibly argued as an outcome variable itself. Since the choice of how much water to consume is endogenous to household decisions, selection, in this case, would be on unobservable variables, not observables, which dramatically limits our ability to match beneficiaries (the matching and simple regression tools rely on the assumption that selection is on observables).

The third alternative eliminates some of these problems—average consumption at the neighborhood level is arguably exogenous to any single household's decision process. It is also easily constructed using survey data. However, neighborhood averages can continue to be plagued by problems of unobserved neighborhood effects (a source of endogeneity) and reflect a mix of supply and demand sides of the water equation. Both options could be estimated using household metering data, although not all households have meters and according to both utilities and customers, the meters may be unreliable, as many are old or poorly calibrated. Accordingly, we propose to supplement household metering data with household surveys.

MORUWASA never attempted a NRW study and the SI team discovered that tight DMAs do not exist for its service areas. Thus, measurement of access in Morogoro will have to rely on a combination of the alternatives discussed above. We will assess, using the data from Dar, the concordance between use of meter-measured access and these alternatives.

Limitations of the NPS data

Between August 25, 2010 and October 3, 2010, MCA-T contracted NBS to collect survey data in 80 clusters in Dar es Salaam, Morogoro and Pwani. These data are referred to as the MCA-T data. The MCA-T data were expected to serve as a baseline for MCC's Water Sector Impact Evaluation in Dar es Salaam and Morogoro. Enumeration of the MCA-T data was part of training for the second round NPS, and the survey instruments were the same as those used in the second round of the NPS. The 80 MCA-T clusters are, thus, in addition to those forming a regular part of the NPS. Like the core NPS sample, it relies on two-stage cluster sampling with 8 households per cluster.

The SI team conducted an assessment of the MCA-T data and concludes that the data are not suited for use as a baseline in a robust evaluation of Water Sector Investment impacts. This conclusion is based on three factors: (i) the timing of the survey; (ii) its coverage, including sample size and geographical focus; and (iii) the questions and information available from the survey. Annex II provides detailed information on the assessment.

Evaluation Design

Introduction

To identify the impact of the water project, we wish to compare the outcomes of individuals who have received increased availability of water against the counterfactual: the outcomes for these same individuals, *if they had not received increased availability through the water program*. Since it is not possible to directly observe the counterfactual, we need a mechanism to estimate it with as little bias as possible. The ideal method is to randomly assign participation among a sample of potential participants, creating a treatment and control group. Through random assignment, the treatment and control groups, on average, are expected to be equivalent along all characteristics affecting the outcome of interest. Hence, in the absence of the project, both groups would have the same expected outcome and any differences between the two groups after project implementation can be attributed to the project.[11] For the Morogoro and Dar es Salaam water improvements, participation is *not* randomly assigned; in fact it appears as though everyone in both cities will be affected to varying degrees by the change in access to and quality of water. It is not possible to go back and randomize participation retroactively. One means of randomization would be to use flow control valves to vary access randomly throughout the city. For technical, political and ethical reasons, however, such a design is not feasible. Since the ideal experimental design is not feasible, we propose alternative quasi-experimental methods for identifying counterfactual outcomes. Below, we give attention to our preferred methods, including (i) use of a generalized propensity score matching process together with difference in difference techniques and (ii) a structural instrumental variables approach, also with difference in difference techniques. The quantitative approaches complement each other and, since data needs are identical for (i) and (ii), the marginal cost of triangulating the results through a second method is created by costs of analysis, which are relatively low (compared to data collection). We also present an overview of qualitative approaches and make recommendations relative to a preferred mix of quantitative and qualitative approaches.

It is important to realize that the impact of the various sources of bias complicating construction of a counterfactual will vary by outcome and other factors. That is, the bias could exist, but its relevance is an empirical issue. In some cases it will be very important and in others it may be less so. In cases where it is less so, precise estimates of impact can be garnered from relatively simple methods. For example, changes in time spent hauling water are likely to be largely due to changes in availability, and the absence of intervening factors (sources of potential bias) makes it relatively easy to construct a counterfactual. We will take efforts to use knowledge of the relationship between the treatment and the outcome to judge the degree of bias and present alternative estimates of impact using the different techniques.

[11] Assuming a well-run experiment without spillovers, differential attrition, Hawthorne effects, etc.

Below we describe our recommended[12] design options for the MCA-T impact evaluation. We stress that none of the methods here should be seen as competing. In fact, data needs for applying each of these options are similar; if we collect data for a GPSM we also have data for the IV approach. As noted by several authors, matching methods and regression-based model adjustments should also not be seen as competing but rather as complements. Much research (e.g. Rubin and Thomas 2000; Ho et al. 2007; Galiani et al. 2005) has shown that the best approach is to combine multiple methods. For example, regression analysis (such as IV) can be conducted on matched samples. Selecting matched samples reduces bias due to covariate differences, and regression analysis on those matched samples can adjust for small remaining differences and lead to increased efficiency (Stuart and Rubin 2007). As noted above, the magnitude of bias is an empirical question and for those outcomes where bias is expected to be relatively small, we will rely on more precise and simpler techniques. We strongly urge that the quantitative household data analysis be complemented with qualitative techniques to fill gaps in the analysis and deepen understanding of the mechanisms through which impacts are realized.

Option 1: Generalized Propensity Score Matching
Since virtually all households in the population will be affected by increased water availability induced by the MCA-T investments, the evaluation approach is conceptually different from a binary treatment approach. The impact of the water investments can be analyzed using tools of the recently developed literature on continuous treatment (Hirano and Imbens 2004; Bia and Mattei 2008).

The GPSM approach assumes that conditional on the vector of baseline characteristics, X, the expected outcomes of the treatment and comparison groups are independent of the assignment. In other words, we assume that controlling for X, selection bias is removed and the comparison group becomes a valid counterfactual. Of course, the potential exists that unobserved variables will differ across the treatment and comparison group, thus violating this Conditional Independence Assumption. However, by including as many relevant predictors of participation as possible in the calculation of the propensity score, we can minimize the likelihood of unobserved variables creating selection bias.

Households in the Dar and Morogoro service areas will benefit from improvements as there is expected to be increased availability of water in both systems. The treatment can be interpreted as the change in availability of water. Following Hirano and Imbens (2004), we let T stand for the treatment (change in water availability), X for a set of covariates, and R for the Generalized Propensity Score (GPS)[13]. If $T_i|X_i \sim N(\beta_0 + \beta_1' X_i, \sigma^2)$, the GPS can be estimated as:

$$(2)\ \ \hat{R}_i = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} exp\left(-\frac{1}{2\hat{\sigma}^2}(T_i - \hat{\beta}_0 - \beta_1' X_i)^2\right)$$

This generalized propensity score is analogous to the standard propensity score when treatment is continuous; households/observations can be matched based on the score.

[12] Description of an alternative option, regression analysis, in included in Annex V.

Implementation of the GPS method consists of three steps (Bia and Mattei, 2008). In the first step, we estimate the score r(t, x). In the second step, we estimate the conditional expectation of the outcome as a function of two scalar variables, the treatment level T and the GPS R: β(t, r) = E (Y | T = t,R = r). In the third step, we estimate the dose–response function, μ(t) = E[β{t, r(t,X)}], t ∈ T, by averaging the estimated conditional expectation, $\hat{\beta}${t, r(t,X)}, over the GPS at each level of the treatment.

In practice, we estimate

$$(3)\ E[Y_i|T_i, R_i] = \alpha_0 + \alpha_1 T_i + \alpha_2 T_i^2 + \ \alpha_3 \hat{R}_i + \alpha_4 \hat{R}_i^{\ 2} + \alpha_5 T_i \hat{R}_i$$

As a result,

$$(4)\ \widehat{E}\ [Y_i|T_i, R_i] = \frac{1}{N}\sum_{i=1}^{N}(\hat{\alpha}_0 + \hat{\alpha}_1 t + \hat{\alpha}_2 t^2 + \hat{\alpha}_3 \widehat{r_i}(t, X_i) + \hat{\alpha}_4 \widehat{r_i}(t, X_i)^2 + \hat{\alpha}_5 t\widehat{r_i}(t, X_i))$$

The Average Dose-Response Function (ADRF) is then obtained by estimating E [Y |T_i , R_i] for every value of t (which implies re-estimating *r* in each stage). To test for the effect being zero, we conduct a joint significance test of the estimated α variables in (4). Under the alternative hypothesis (that at least one of the coefficients is not zero), the F statistic has a non-central F distribution. A non-central F distribution is the ratio of a non-central chi-squared and a (conventional) chi-squared random variable.

The continuous treatment approach embodies the central idea that there is only one "group", because all households are treated (there is no control group), and what varies is the intensity of the treatment assigned to each household. Hypothesis testing is complicated by the need to test joint hypotheses (rather than using t statistics, we need F statistics). Software for estimating these effects is directly available in Stata (Bai and Mattei, 2007).

Option 2: Instrumental Variables Regression Approach

A further option is to use an instrumental variables approach. This approach helps overcome weaknesses of DD and other methods that fail to control for sources of selection bias that change over time; they also relax the conditional independence assumption underlying the GPSM procedures. Start with a linear treatment model:

$$(5)\ Y_i = \beta X_i + \ \alpha W_i + \varepsilon_i$$

Where Y is the outcome, X represents a vector of covariates, W is water consumption (possibly a vector representing quality and quantity dimensions) and α is the effect of increased/improved consumption of water on the outcomes. If treatment is randomly assigned, then selection bias is not a problem, but, as noted above endogeneity may exist. Endogeneity means that cov(W,ε)≠0, which violates one of the key assumptions in OLS which is generally used to estimate equation 5. Endogeneity may result from program placement (e.g. the water system is placed in more favorable places) or from unobserved individual heterogeneity resulting in individual decisions about W.

Instrumental variables estimation helps clean up the correlation between W and ε. To do so, we must find an instrumental variable (Z) that is correlated with W (cov(Z,W) ≠0) and is uncorrelated with ε (cov(Z, ε)=0. That is, Z affects W, and only affects the outcome (Y) through its impact on W. The instrument helps "purge" the correlation between W and the outcome error. Several means exist for estimating the IV (see Cameron and Trivedi 2005) with the most intuitive being 2-stage least squares (2SLS). Start with a first-stage regression:

$$(6)\ W_i = \gamma Z_i + \varphi X_i + u_i$$

The prediction from this regression $\widehat{W}$ only reflects exogenous variation in the treatment. In 2SLS, $\widehat{W}$ is substituted back into equation 5, which can then be estimated by OLS, since, by construction, cov($\widehat{W}$, ε)=0. The second-stage regression estimates our treatment effect (α):

$$(7)\ Y_i = \beta X_i + \alpha(\hat{\gamma} Z_i + \hat{\varphi} X_i) + \varepsilon_i$$

Since $\hat{\gamma} Z_i + \hat{\varphi} X_i$ is, by construction, uncorrelated with the error term, the problem of endogeneity is eliminated.

The specific variables and form of equation 7 will depend on the choice of Y and will be guided by economic theory. For example, if Y_i is taken to be time spent gathering water by individual i, we would employ a model of time allocation, and X would include the appropriate variables in their appropriate form. A model of water-related illnesses would imply a different functional relationship and different regressors (the dependent variable would be individual-level presence of a water-related illness in the reference period, severity and duration, if information on the latter is available from the survey).

The IV approach has several appealing properties. First, and most important, a regression-based framework allows examination of behavioral relationships. While the overall focus of the analysis is a clean measurement of impact; regression-based analysis provides insights into why the intervention had or failed to have certain impacts. For example, we can examine the functional relationship between the covariates and intervening variables, such as source of water, decisions about consumption of water, etc. Evaluations of large-scale interventions in the water sector have often failed to find significant impacts; use of regression-based causal models allows the analyst to examine different pathways of impacts.

Second, an IV approach reduces or eliminates the problem of selection on unobservables. In fact, when combined with a panel data approach (discussed in more detail below), IV estimation permits a nuanced view of unobserved heterogeneity—the panel approach eliminates time-invariant unobserved heterogeneity while the IV cleans up time-varying sources. Third, IV estimation can be combined with matched data which reduces problems associated with imbalance between covariates in different treatment groups.

Three main criticisms of IV approaches have emerged: (i) assumption of a particular form of the relationship between the treatment and the outcome; (ii) weak instruments; and (iii) the challenge of heterogeneous treatment effects. We will address (i) by estimating alternative specifications and evaluating the robustness of findings to the different forms. This is one component of the triangulation we mention above. The problem of weak instruments depends on the data. When the instrument is correlated with unobserved characteristics affecting the outcome a bias will result. If the instruments are only weakly correlated with W, the standard errors of the estimates are likely to increase. We plan to address the issue of weak instruments in two ways. First, our main instrument is water supply (defined above), which is clearly related to W but only related to Y through its relationship with W. Second, we will remove the correlation mentioned above by employing first differences (with panel data). The differencing will remove the influence of time-invariant unobserved factors; the IV estimation will address the remaining time-varying unobservable factors.

It is widely recognized that IV methods are unable to capture either the average treatment effect (ATE) or the average treatment effect on the treated (ATET) when treatment effects are heterogeneous. Intuitively, this is because only a subset of the population is affected by any particular instrumental variable and the model will not capture effects for those people whose treatments are not affected by the instrument. Angrist, Imbens and Rubin (1996) call the population of potential beneficiaries whose decision to participate is affected by the instrument "compliers". Without making too many assumptions, IV methods can be relied on to capture the effect of treatment on compliers (Angrist 2012). The average effect for this group is called a local average treatment effect (LATE), first discussed by Imbens and Angrist (1994). Fortunately in the case of a continuous treatment and the widespread influence of the MCA-T water projects, the LATE is, in fact, the parameter of interest. To the degree that availability of water affects decision making, everyone in the population will be affected by the IV (here, the instrument will be change in availability of water) and IV estimation will yield the relevant policy parameters: impacts of water availability on outcomes of interest for compliers.

Hoderlein and Sasaki (2011) combine and extend two literatures: instrumental variables estimates of continuous treatment effects and estimation of causal structural models. They provide a framework in which causal effects of continuous variables, or treatments, can be understood when there is endogenous selection of the continuous treatment intensity. They build on literature by Heckman and Vytlacil (2007), which established a causal link between an endogenously determined binary treatment variable (X) and an outcome (Y). Heckman and Vytlacil derive a useful parameter (the marginal treatment effect, or MTE), or the derivative of the outcome with respect to the probability to be treated. This MTE is identified with a local instrumental variable with continuous instruments (Z). Hoderlein and Sasaki extend the Heckman and Vytlacil results to the case where the endogenous treatment (X) is continuous and derive a useful analog to the MTE, called the local average structural derivative (LASD), which can be used to measure aggregate causal effects. Estimating the LASD involves a straightforward application of instrumental variables methods.

Summary

Our recommended quantitative methods are a generalized propensity matching (GPSM) procedure and an instrumental variables (IV) structural model. Since virtually all households in the population will be affected by increased water availability induced by the MCA-T investments, the evaluation approach is conceptually different from a binary treatment approach, as explained above. The impact of the water investments can be analyzed using tools of the recently developed literature on continuous treatment (Hirano and Imbens 2004; Bia and Mattei 2008). The GPSM approach assumes that conditional on the vector of baseline characteristics, X, the expected outcomes of the treatment and comparison groups are independent of the assignment. In other words, we assume that controlling for X, selection bias is removed and the comparison group becomes a valid counterfactual. The continuous treatment approach embodies the central idea that there is only one "group", because all households are treated (there is no control group), and what varies is the intensity of the treatment assigned to each household.

Of course, the potential exists that unobserved variables will differ across the treatment and comparison group. To account for this, we also plan to incorporate an instrumental variables approach. The IV approach provides an instrument, or a variable that is correlated with water consumption, for example, but is not correlated with other unobserved variables. By including this variable as an instrument in a standard regression analysis, we can more clearly evaluate the impact of interventions on water consumption across the sample. This approach helps overcome weaknesses of DD and other methods that fail to control for sources of selection bias that change over time; they also relax the conditional independence assumption underlying the GPSM procedures.

Both techniques will be applied to data from the panel of households sampled in the baseline and follow-up surveys. Panel data allow us to construct the variables of interest—changes in availability of water and changes in outcomes. They also allow us to eliminate the effects of time-invariant observable and unobservable variables. This elimination allows us to reduce bias associated with these factors and improves the reliability of either causal estimate.

In addition to these causal estimates, we intend to assemble time series data from other sources, including health indicators from the Ministry of Health's Health Information System and its District Medical Offices (DMOs), and National Bureau of Statistics (NBS) data on business activities and start-ups in the cities. The data collection firm will be tasked with gathering data from these sources. These data do not contain the main outcomes of interest, and there is no valid counterfactual at this aggregated district level, so these data will be used to create descriptive indicators of changes before and after project start-up in the health and economic context in which the interventions are taking place.

Threats

Intervening variables

The biggest threat to achieving the expected impacts is the necessity of completion of the transmission main between the Lower Ruvu treatment plant and the University reservoir. Without this transmission main, the treatment plant will not begin operation. Other intervening variables have strong potential to sever the link between increased water availability and quality and positive outcomes shown in Table 1. These variables include differences in the quality of water displaced by the investments (e.g. water that was consumed from different sources), and quality changes in storage and hauling. Several experts have noted that water storage, which is common in Tanzania, may lead to contamination of clean water already extracted from pipes. This is an especially acute problem for households that haul water from other sources (neighbors, kiosks, etc.). As a result of this potential contamination, the improved supply of water may not lead to desired outcomes like improved health and reduced school absenteeism. SI recommends testing water, as discussed below, for fecal coliform at its source and when it is consumed to identify the significance of this potential problem; these measurements will be conducted for all surveyed households[14].

Measurement

Measurement of availability of water is a major challenge. Each of the alternatives discussed above has its shortcomings, and we expect water availability to be measured with some error. The flow meters (at the DMAs) will need to be tested and inspected regularly to ensure measurement integrity. In Morogoro, bulk water meters are being installed to measure water production but DMAs do not exist, and we will not have instrumentation to measure water supply and changes in it at the household level. We will have to rely on household measurement of access supplemented with area agent options (the alternatives described above), and uptake averaged over neighborhoods. We will use the Dar es Salaam data to analyze the correlation between supply measurements from metered areas and corresponding neighborhood averages; information on this correlation will help confirm the use of this averaging method in Morogoro. Since water comes into the household irregularly and through many sources, measuring water consumption/use is not straightforward. It is expensive and intrusive to measure use through observation, so the best alternative is to rely on own reporting or recall. SI recommends an overhaul of the NPS survey instrument to better measure volume consumed (see Annex II for a discussion).

Attrition

With data being collected in post-program follow-ups, attrition, or the inability to collect follow-up data from households selected for panel data[15] collection, could be a significant threat to internal validity. Indeed, working in urban and peri-urban environments, our team expects a number of households involved in the study will migrate for economic reasons during the life of the evaluation. If such attrition is random, then the threat is minimal, simply reducing power. This can easily be corrected by including a buffer in the required sample size to account for non-response. Unfortunately, attrition is rarely random. It is often correlated with treatment, meaning that households experiencing more or less availability of water might be more likely to leave the sample than the other. For example, households receiving only minor amounts of additional water availability may be more likely to move in search of availability than those that receive substantial amounts. This creates an imbalance or selection bias between varying by treatment intensity which can lead to biased impact estimates.

The most important mitigation strategy is collecting good location information. Particularly in urban areas, most households in Tanzania have access to cell phones. This will be a primary source of contact assisting us in the location of targeted respondents, but we will also collect more detailed location information. This will include information about the individual's current location, such as address, email, and phone numbers, as well as information for people who would know how to locate them if they were to move, such as their parents, friends, or other family members. By collecting such detailed location information, we reduce the likelihood that attrition will occur simply because an individual will have moved.

[14] The team is currently exploring costs of water quality testing in Tanzania.
[15] Collection of panel data is highly recommended as it can be used to eliminate biases from unobserved individual and household-level effects and, thus, improves the precision and reliability of the estimates.

Nevertheless, we anticipate that inevitably some degree of attrition will still occur. The evaluation team will take the following steps to measure and mitigate the effects of attrition:

1. Budget for Additional Follow-ups: Within the data collection budget, a small buffer should be included to fund additional tracking of non-responders.
2. Track a Sample of Non-Responders: Depending on the level of attrition, the evaluation team will select a random sample of the non-responders for additional follow-up. This additional tracking is resource intensive as it often involves talking to multiple people in order to locate one non-respondent.
3. Analyze Attrition Data: The data collected from the sample of non-responders will be analyzed for patterns of non-response, including in terms of baseline characteristics, assignment, and outcomes correlated with non-response.
4. Bound Potential Attrition Bias: From the sample of tracked non-responders, the evaluation team can extrapolate the outcomes of the population of non-responders providing a revised estimate of program impact. This analysis can be done under a few sets of assumptions, including best and worst-case scenarios that provide upper and lower bounds of program impact.

DATA

Data Needs

Household Survey

Data needs vary according to the indicator (Table 5). The primary outcomes of interest are individual- and household-level changes in key outcomes, and data must come from a household survey[16]. Since we are interested in changes, we need individual-level observations before and after the facilities are in operation, or at least two rounds of data collection. Ideally, the MCA-T survey (described in Annex II) would have sufficed as a baseline; the survey questionnaire is well-done and its breadth would allow a variety of analysis. Unfortunately, the survey is not suitable due to issues detailed in the Annex. Thus, we recommend commissioning a new household survey for Dar es Salaam. One round of the survey will take place prior to start-up of the system expansion (and completion of the transmission main) and the follow-up will occur approximately one year after full operation of the system. This time frame is selected primarily on the basis of measuring outcomes and short term objectives. Longer term objectives and the goal indicator may take longer to materialize. However, since we do not expect large effects in these outcomes, they would require prohibitively large sample sizes to measure. Moreover, longer follow-ups will increase costs (through respondent tracking) and introduce additional intervening variables (including proposed expansion of the Upper Ruvu treatment plant). Accordingly, we will rely on secondary data to measure these outcomes, and triangulate with one-year follow-up data. Shorter follow-up periods would introduce seasonality bias.

In Morogoro, the survey timing is more compressed, as some of the MCA-T investments will begin operation later this year. As a result of this compression, data collection recommendations are different for Morogoro (see below). Social Impact strongly recommends use of electronic data entry in the field. This technique allows for real-time identification of outliers and consistency checks. It will not add appreciably to enumeration costs.

[16] Without household-level data, we will be constrained to do before-after comparisons with existing data (possibly the 2010 DHS and its follow-up, expected in 2015).

Water Quality Data

Three indicators of water quality are proposed; turbidity expressed in nephelometric turbidity units (NTU), a bacteriological indicator expressed as Coliform Microbial Density (CMD) per 100 milliliters, and a chemical parameter, Free Chlorine Residual (FCR). These measurements require specialized instrumentation and procedures and will carried out by an appropriately qualified and experienced local contractor engaged by SI. Further details concerning the design of water quality data collection for the IE are presented below.

Turbidity*:* It is proposed that turbidity (NTU) be measured at the outlet of treatment plants, outlets of main service reservoirs at supply zone level and at a sample of filling stations, water kiosks and household taps. It should be noted that the Tanzania drinking water quality standards do not specify a reference value for NTU. Therefore, actual values will be compared with the test results from other historical and contemporary sources in Tanzania.

Bacteriological: It is proposed that coliform bacteria, expressed as Coliform Microbial Density (CMD), be measured both from the tap and in storage containers of the household sample as well as a sample of water tankers and water containers in all supply zones in Dar es Salaam and Morogoro. The reference value for human drinking water is 0 coliform colonies per 100 ml. The methodology for CMD will be to compare test results to the reference value, expressed as a percentage of samples not meeting the reference criteria.

Chemical: Free Chlorine Residual (FCR) is the proposed chemical indicator of water quality. It is proposed that FCR be measured at the outlet to the treatment plants, at outlets to main reservoirs in Dar es Salaam and Morogoro, and at a sample of filling stations, water kiosks and taps at households and institutions. It should be noted that the Tanzania drinking water quality standards do not include a reference value for FCR. However, the range of values for FCR is typically between 0 and 5 mg/L. The WHO guideline value for FCR in drinking water is 5 mg/L. The proposed methodology for use with FCR will be to compare the concentration of chlorine at the outlet of the treatment plants with the concentration at key points within in piped network, i.e. outlets to main service reservoirs, and a sample of filling stations, water kiosks and taps at households and institutions, expressed as a percentage of test results which do not meet the reference criteria.

Assumptions: Where possible and where reliable and timely data are available, data from the two water utilities (DAWASCO and MORUWASA) on NTU and FCR will be used to establish baseline values for turbidity and chlorine concentrations. Turbidity and chlorine concentrations are typically regularly measured at water treatment plants as a standard procedure in urban water supplies, although at the time of this writing, it is not yet known if this will be the case at the Lower Ruvu and the two treatment plants in Morogoro. If reliable and timely data on turbidity and chlorine concentrations cannot be obtained from DAWASCO and MORUWASA, it will therefore be necessary for SI to engage a suitably qualified and experienced contractor to collect and analyze the required samples.

Bulk Meter Readings

DAWASA has installed a system of approximately[17] 75 bulk meters, which are intended to be read twice per month. Given the regular power interruptions, many of the bulk meters are connected to UPS devices, although DAWASCO engineers reported that the power was often out for longer than the UPS could support. The team visited one bulk meter, which was not functioning due to power outage. Through the MCC NRW component, DAWASA was also provided with data logging equipment, although it appears that they are not currently in use. If these meters and logging equipment can be reliably used, they could provide a good source of data on water availability through each of the 23 DMAs in Dar. Since this data serves as our preferred measurement of treatment level, the team will need to conduct an additional assessment of the current status and availability of the bulk metering and logging equipment. We do not expect the need to purchase additional equipment (if the current and NRW equipment is operational), although the team may need support from local staff to collect meter readings. If the equipment is not available or functional, we will rely on household level responses on water availability, as described above.

Additional data sources

The household data will be complemented with alternative sources. While these alternative sources will not allow us to estimate causal effects (no convincing counterfactual) they will help corroborate survey findings. They will also assist in providing nuance to the quantitative evidence. District medical officers (DMOs) are responsible for collecting data on incidence of diarrheal and other diseases at the facility level for all facilities in their district (roughly 180 facilities per district). These data could be used to create maps of incidence of diseases throughout the city, and changes over time could be monitored. The main challenges associated with these data are uneven reporting (fewer than 60% of facilities in Dar report regularly). Several alternative data sources (e.g. the DHS, the GIZ study on urban water poverty, the HBS and the NPS) have information on sources of drinking water in urban areas. As available, summaries of these data will be used to put findings from our household survey in perspective.

[17] A precise number was unavailable from DAWASCO engineers during the site visit.

Table 5: Indicators

Result	Indicators	Level of Measurement	Modification to NPS Instrument?	Comments
Outcome				
Water service coverage	# of domestic / non-domestic customers	Utility	N/A	Compiled from utility administrative records
	% active customers	Utility	N/A	Compiled from utility administrative records
	% of households with access to improved water supply	HH	Yes: Slight modification	Proportion whose main source of drinking water is a household connection (piped), public standpipe, borehole, protected dug well, or rainwater collection (DHS)
Water service quality	Average hours of service per day	Utility and HH	No: NPS asks for average hours per day and days per week	Household measurement is critical for estimating treatment. See discussion above for suggestions on improving and triangulating NPS data.
Water quality	Nephelometric turbidity units (NTU)	Utility and HH (pipe)	Yes: Protocol for water testing	Requires separate water quality testing at the plant and point of use
	Coliform Microbial Density	Utility and HH (pipe and storage)	Yes: Protocol for water testing	Requires separate water quality testing at the plant, point of use (pipe), and storage
	Free Chlorine Residual	Utility and HH (pipe and storage)	Yes: Protocol for water testing	Requires separate water quality testing at the plant, point of use (pipe), and storage
Water consumption	Volume of commercial / residential water consumption	Utility, Business, and HH	Yes: Requires modification to measure volume consumed	Utility level data compiled from administrative records. Business and HH data will be triangulated through administrative records and surveys to account for other sources
Water expenditures	Average daily water expenditure; share of water expenditures in total household expenditures	HH	Yes: More focus on water expenditures needed	NPS questionnaires ask about water expenses, but a better focus on these expenses is needed
	Average daily water sales	HH and business	Yes	Households with new/additional access may sell to community members. Business data through qualitative study
Water security	Perceived water security[18]	HH	Yes	Section I of questionnaire would have to be modified
Short-term Objectives				
Water-borne	% of population with	Individual –	Yes: survey only asks for	Low prevalence in >5 and during dry season will make

[18] Water security is defined as a household's level of access to sufficient, potable, continuously available wáter and the relationship of that wáter access to the health and productivity of that household's inhabitants.

disease related morbidity	diarrhea in the last 2 weeks	including gender and age	<5	identifying impact difficult
Human capital accumulation	Average hours worked last week	Individual—including gender and age	No	This can be valued using wage/earnings information from the survey
	Percentage of children who missed any school in the last 4 weeks	Individual—including gender and age	Yes: change recall from 2 to 4 weeks	Expect seasonal variations
	Average time spent fetching water from home in last week	Individual—including gender and age	Yes: include information on individual responsible	We will also estimate value of time using wage rates (if adults) or some fraction of wages, if children
Medium-term Objectives				
Mortality	<5 and adult mortality	National	N/A	From DHS. No counterfactual and small anticipated effect size, so attribution will be difficult
Economic activity	Average current value of HH / commercial assets	HH, business	No: HH asset register	NPS must be supplemented by business survey and qualitative data
Compact Goal				
Poverty Reduction	Average annual household income per capita	HH	N/A	Collected from HBS

Sampling

Sample Size

The SI team conducted an analysis of the minimum sample required to produce reliable statistical estimates of key outcomes.[19] We assume a clustered, quasi-randomized evaluation design with a continuous treatment whose magnitude varies by cluster and by household within clusters. We also assume that data collection will occur before and after inauguration of the projects. Because of the differences in timing of project completion, we are treating Dar es Salaam and Morogoro as distinct entities. As a result, we are computing separate sample size requirements for each. Also, in order to obtain data representative of an entire year, we recommend collecting one half of the sample during the rainy season and one-half during the dry season[20]; findings will not be representative of rainy or dry season, but will be representative of the entire year.

The purpose of the sample size estimates is to determine the minimum impact that can be detected for a given sample size. Sample size calculation includes the number of clusters in the sample and the number of households within each cluster. If the measured impact of the treatment is at least as large this minimum impact, we will be able to detect it 80 percent of the time with a given sample size. If the treatment impact is less than the calculated minimum impact, however, we are less likely to detect it. In determining the sample size, we used several alternative outcome variables: household expenditures on water, time spent collecting water (wet season and dry season), household exposure to serious water shortages, and indicators of diarrheal disease among children. These indicators were chosen because they coincide with the project logic as spelled out in the M&E plan and information on them was available from the recent NPS MCA-T household survey.

Results from the power analyses are shown in detail in Appendix III. In sum, a sample size of 2500 households from each site (Dar es Salaam and Morogoro) would provide sufficient power to detect statistically significant changes in three important outcomes: household expenditures on water, time necessary to haul water in the rainy season, and changes in exposure to major water shocks. Differences in time spent hauling water during the dry season (more time is spent in the dry season) would be harder to detect with a sample size of 2500, but we might be able to detect the difference depending on the magnitude of the effects among other parameters. Differences in rates of children's diarrhea would be extremely difficult to detect, requiring a sample size of more than 200,000. The sample of 2500 households in each site will be split evenly across the periods of time representing dry and rainy seasons (so that 1250 will be interviewed in each season, in each site).

Timing

The two major factors that determine the proposed timing for data collection include seasonal variation and variation in the projected completion dates for the improved works in each site. Seasonal variations in outcomes are the norm as the areas under study undergo a long rainy season from March to May and a shorter season from November to December. Negative health outcomes are closely correlated with the rainy seasons, and water from alternative sources is also more abundant (and likely to be less expensive) during this period. We would need to time the survey enumeration so that the measured (particularly outcome) variables accurately reflect conditions during the wet and dry seasons. In each season we would also ask respondents of the household surveys and participants in the qualitative research groups to self report water sources and gathering times for the other season as well. For questions reflecting a long period of time (such as information over the past year), this timing is not critical. But for more short-term recall periods,

[19] Detailed outline in Annex IV

[20] The recall responses from the MCA-T NPS oversample show that there is pronounced seasonality in water sources, travel time to the sources, and other factors related to water purchases. This sample was conducted in August-September, a dry-season period in both cities, but asked recall questions about water sources and time to the sources for both rainy and dry seasons. To obtain representativeness of the rainy and dry season, the sample sizes would need to be roughly doubled.

we need representativeness for both periods. Thus, we would stratify by season, with the survey being timed to get representativeness across rainy and dry seasons in both sites.

In Dar es Salaam, the earliest the MCA-T expansion is likely to be functioning is in May 2013, although possibly much later. First round data collection needs to reflect this timing; the baseline surveying needs to be complete before operation. The follow-up survey will be conducted approximately 1 year after the baseline survey, respecting the seasonal patterns described above. A one-year window will allow for most impacts to be realized, but longer-term outcomes such as increased business activity and higher household incomes due to increased productivity will likely not be fully realized during this time. Changes in longer-term outcomes can be measured using secondary and other data sources, and establishing a causal link between the water investments and longer-term outcomes will be difficult using the methods we propose. If there are significant delays in construction of the transmission main from the Lower Ruvu treatment plant to the University reservoirs in Dar, this approach can simply be delayed until the treatment plan is scheduled to come online, as it is somewhat independent of the approach in Morogoro[21].

In Morogoro, some of the improvements will be in operation in July/August 2012, and given the contracting time required by MCA-T, it will be impossible[22] to field a first-round survey prior to this time. We recommend using the (limited) MCA-T 2010 survey augmented with the 2010/11 NPS as a baseline. The follow-up will be conducted approximately 1 year following full operation of all the components. The sample design of the follow-up survey will be modified to reflect the sampling of the MCA-T and NPS samples; the sample size will be expanded to reflect the geographic spread of impact and seasonality.

An alternative, and preferred, approach exploits the staged increases in water availability through the phased completion of the works in Morogoro. In this approach, we would utilize the existing NPS data as a baseline, supplemented by a midline survey at the end of 2012, prior to completion of the new treatment plant. This survey serves as a follow-up for the NPS baseline (measuring change from the first set of completed works) as well as a baseline for a follow-up survey collected at the end of 2013.

Qualitative Data

Qualitative data will be collected and analyzed to provide additional insights into the processes and subjective factors which are expected to influence the outcomes and impacts of improved water supplies and which do not lend themselves to quantitative methods. Qualitative data will be generated through semi-structured interviews with key informants, focus group discussions (FGDs) with important target groups, direct observation and case studies in Dar es Salaam and Morogoro.

Semi-structured interviews will be administered to the following key informants:

1. MCA-T technical and evaluation staff. Interviews will focus on perceived program success and challenges in implementing programs.
2. DAWASA/DAWASCO/MORUWASA management and technical staff. Interviews will center on improvements in collection rates, perceived changes in customer satisfaction, evidence of improvements in well-being, and business activities.
3. Business owners, including formal and informal water vendors. Discussions will emphasize changes in business levels, change in supply, demand, and price for water, changes in water availability, and new business opportunities that result from the improvements. It is important to note that changes in demand for water from vendors or in business revenues do not necessarily suggest a negative program impact, and could be indicative of changes in the efficiency with which public and private resources are used.

[21] Although implementing both approaches simultaneously would likely generate efficiencies in travel, data collection, and analysis.
[22] Unless perhaps contracting for data collection can be done through SI.

Focus group discussions (FGDs) will be used to elicit information from the groups listed below. While there is some degree of sensitivity in eliciting information from some of the groups listed, such as informal business owners or young children/minors, the qualitative research methods will be designed to protect subjects and guarantee confidentiality in order to maintain the integrity of the data collection among these groups while minimizing non-response. Other groups may be identified and added as the study progresses.

1. Water kiosk operators and customers.
2. Female water users (DAWASCO/MORUWASA customers and others).
3. Child water carriers (ranging from 6-18 years of age).
4. Business groups, especially small-scale and informal type businesses most affected by inadequate access to water.
5. Representatives of schools and health facilities.

The qualitative studies and data will be designed to provide information in the following areas:

1. Adequacy of water services and how adequacy has changed over time.
2. Satisfaction with water services and changes over time.
3. Perceptions of water quality and how it has changed, emphasizing changes in water purification practices and changes in perceptions about illnesses and their incidence and severity.
4. Perceptions of water agencies' performance and responsiveness.
5. Impacts of increased time availability due to changes in time spent hauling water; impacts on school attendance and time spent working.
6. Changes in cost of obtaining water for drinking and other domestic uses.
7. Perceptions of recent changes in water services.
8. Recommendations for improvements in water services.

The baseline qualitative study will take place as close to the time the planned improvements are implemented as possible and will be repeated approximately one year following the baseline data collection with the same key informants and, where possible, FGD members.

In order to provide additional perspective, insight and detail to the evaluation, it is also proposed that a number of case studies be prepared on the basis of the proposed qualitative methods. These case studies will include, but not be limited to, the following groups:

1. Female water kiosk operators
2. Water tanker owner/operators
3. Pushcart water delivery operators
4. Female heads of household in low-income areas
5. Owners/managers of new water-related businesses, e.g. households that have water to sell to neighbors

Table 6 (below) presents possible subjects to be explored with the above groups:

Table 6: Proposed Content of Case Studies by Subject Group

Group	Key Content
Female water kiosk operators	- Daily activity diary - Reliability and adequacy of water services (recent trends) - Water quality - Revenue and income - Relationship with utility - Work-related issues/solutions
Water tanker owners/operators	▪ Reliability and adequacy of water services (recent trends) ▪ Water quality

	▪ Customer profile ▪ Revenue and income ▪ Relationship with utility ▪ Business-related trends, issues and solutions
Pushcart water delivery operators	- Daily activity diary - Reliability and adequacy of water services (recent trends) - Water quality - Revenue and income - Relationship with utility - Work-related issues/solutions
Female heads of household in low-income areas	- Daily activity diary/Time spent collecting water - Reliability and adequacy of water services (recent trends) - Water quality - Water consumption - Water expenditure - Water-related income-generating activities
Owners/managers of new water-related businesses	▪ Reliability and adequacy of water services (recent trends) ▪ Customer profile ▪ Revenue and income/profit ▪ Relationship with utility ▪ Business-related trends, issues and solutions

The case studies will consist of in-depth interviews, direct observation and, where relevant, examination of pertinent documents/records. Subjects will be chosen in areas that are likely to benefit from the planned improvements in water services, and to the extent possible, will be revisited after the planned improvements have been implemented. Where deemed necessary, the case studies may be undertaken over several days or at intervals over weeks or months to capture seasonal and other variations.

Description of Key Variables

Survey instruments

Our main household survey instrument will be close to the NPS instruments with modifications to allow measurement of key variables. The NPS is comprehensive, perhaps to a fault as it was reported to take on average more than four hours to complete[23]. Survey costs could be reduced by eliminating some unnecessary (for the purpose of our analysis) sections. We have ordered the questionnaire sections into three tiers: tier 1 is essential for the analysis; tier 2 would add some value, but not essential; tier 3 would be easily deleted. **Tier 1** includes: Sections A-F; Section I (food security—converted to questions about water security); Section J (water and sanitation); Sections K-M (expenditures); Section N (assets); Section S (deaths); Section V (contact information). **Tier 2** includes: Section O (groups—contains information on charitable receipts); Section Q (finance); Section U (anthropometry—if nutrition indicators are used as outcomes). **Tier 3** includes: Section G (subjective welfare); Section H (governance); Section P (credit); Section R (shocks—but the question on water shock should be moved to section J). The agriculture and fisheries questionnaires could be excluded, but the community questionnaire is necessary.

Variables

Outcome variables and how they will be measured are shown previously in Table 2. All of these variables will be collected via the household survey, administrative records, secondary sources, and water quality testing.

Covariates are described in Table 7. Collection of consumption expenditures is a lengthy and relatively expensive undertaking, but we have included them among our tier 1 variables for two reasons. First, changes in expenditure patterns are likely to emerge as more abundant and cheaper

[23] We aim to keep the instrument under 90 minutes

water sources are available and defensive health expenditures decline. Thus, we would like to understand how these treatment-induced changes impact other expenditure patterns in the household. Second, consumption expenditures per capita are widely taken to best reflect household well-being, and we can use the expenditure measure to disaggregate impacts by socio-economic standing (poor, non-poor, well-being quartile, etc.).

Table 7: Covariates in the Models (Matching and Regression)

Variable (all measured in changes)	Level of Measurement	Modification to NPS Instrument?	Comments
Demographics	Household; based on individual characteristics	No	Age/gender structure; educational attainment of adults
Wealth; household assets	Household	No	Use asset index with data from Section N; also housing from Section J
Geographic	Household	Yes: community questionnaire is geared toward rural areas.	Need to geo-reference households in the survey. Need information on neighborhood characteristics, distances to key facilities, employment possibilities, etc.
Consumption expenditures	Household		Inputs into health production; information on changing expenditure patterns following change in water expenditures; used for measuring household well-being
Availability of water	Household	Yes for a truly exogenous measure	See notes above

Other covariates are related to the ability to apply GPS matching; more information on exogenous household conditions allow for enhanced matching. They are also useful in estimating structural models (such as changes in labor supply, time allocation, health). We will specify IV econometric models of each of these to understand how changes in availability (supply) affect water consumption, and how this change affects the outcomes.

IMPLEMENTATION

IRB Requirements

SI has an internal IRB which will be used to approve and review the study before data collection begins. Participation in a local Tanzanian IRB is not required; however, SI will make contact with the IRB based at the University of Dar es Salaam and request a review of the proposed design. Key contacts at the local IRB will also provide guidance on local issues like respondent compensation, survey timing, instrument design, and so forth.

Analysis Plan

Introduction

Our analysis plan follows the logic of the impact of the water investment. Pathways of impact have been clearly spelled out and confounding factors have been identified. We begin the analysis with relatively simple questions: (i) have we adequately accounted for confounding factors? and (ii) have we addressed problems of biases? We will use quantitative techniques to address these questions. We will examine the consistency of findings across out analysis techniques, and will triangulate using qualitative information.

SI recommends two phases of analysis with the quantitative data: analysis upon completion of the first round of the household survey (baseline); and analysis after the second round (follow-up). This phased analysis allows adjustments to be made in the follow-up survey, and helps identify potential confounding factors that can be addressed in the intervening period. The baseline qualitative study will take place as close to the time of the improvements as possible and will be repeated approximately one year later with the same key informants and FGD members. Where possible, the qualitative research will occur prior to the quantitative, which may enable adjustments to quantitative surveys for improved measurement of variables if needed. In fact, Alwang and Gacitua-Mario (2008) recommend iterative integration of quantitative and qualitative analysis; each form of analysis informs the other and the resulting evaluation becomes more robust as a result. Thorough analysis of first-round household data will facilitate this integration. In addition, thorough analysis of the first-round data will help broaden understanding of what works, and why. Frequently, impact analyses focus closely on whether something has impact and the magnitude of the impact (Deaton 2010); we will try to understand the behavioral mechanisms behind the generation of impacts.

Prior to the start of the analysis, we will review the quality of the data, and confirm that data on the key variables are of appropriate quality for purposes of rigorous analysis. Fortunately, if the data collection is implemented using field-based electronic data entry, some dimensions of data quality can be assessed on a real-time basis.

Round One

As noted above, we will review data quality and confirm measurement of key variables. We will also identify potential data errors at this stage and suggest means of addressing any important errors. Following this, we will:

1) <u>Generate summary statistics</u>: outcomes and covariates
    a. Descriptive statistics of outcome variables and covariates
    b. Statistical analysis of outcomes and covariates by gender (for individual-specific outcomes; household headship for household-level outcomes), by household well-being such as poor, non-poor or quintile (indexed by asset index or consumption expenditures per capita), by availability of water. We will conduct and present significance tests of differences in outcomes across covariates and by treatment intensity. These preliminary significance tests will allow us to validate the power

analysis described above to ensure that our sample size provides reliable power. They will also help validate the sample in terms of balance of observations by treatment intensity.

c. Create maps of areas of water availability (e.g., low, medium and high availability), water consumption by source, and levels of water consumption. This would require overlaying water meter information combined with the block-level DAWASCO analysis, and the household data source. Conduct analysis of spatial clustering and identify hot spots, conditional on the availability of appropriate GIS data. Note: the statistical representativeness of the household data will not allow formal hypothesis testing of many of these variables.
d. Analyze findings to identify themes for analysis.

2) Analysis of determinants of outcomes

Regression analysis will be used to explore the relationship between outcome variables and covariates, recognizing that biases due to unobserved individual and household characteristics might exist. These regressions will enable understanding of the relationship between covariates, such as family size and structure among others, and key outcomes. They will also help validate whether the logical hypotheses in the project document are valid. The preliminary analysis will allow exploration of alternative specifications and will thus inform the combined analysis. A key question to be addressed here is whether our measure of water availability is associated with the different outcomes.

a. *Water consumption/expenditures*. We will estimate a demand function for water, either in volume or expenditure form; a demand-consistent functional form will be used (Nauges & Whittington 2010; Whittington, Lauria & Mu 1991). One potential avenue of inquiry is whether source of water (see below) affects demand.
b. *Determinants of water source.* We will estimate a multinomial logit model and examine how household size, time availability, neighborhood characteristics and other factors affect this choice (see Lechner 2002 for some details and an application of matching using a multinomial logit model). This endogenous choice may be used in a 2-stage model of factors affecting demand (see a). We will also present kernel-smoothed densities of household well-being (expenditures per capita) by main source of drinking water.
c. *Water borne diseases.* We will estimate a health production function (Rosenzweig and Schultz 1983; Lee, Rosenzweig & Pitt 1997) to examine the relationship between health inputs and diseases.
d. *Water security.* This will be a reduced-form IV estimate of the determinants of the outcome. Water availability and (endogenous) water sources will be included among the covariates.
e. *Labor force participation.* This will be a standard model of labor-force participation (see Bardasi, et al. 2010 for references), estimated as a probit model.

These studies can form stand-alone analytical papers.

3) Synthesize findings, integrate with qualitative data findings and prepare plans for additional variable needs.

Round Two

1) Generate summary statistics: changes in outcomes and statistical significance of them.

a. First differences of all outcome variables; tests of significant differences.

b. DD approach, where the index is gender (for individual-specific outcomes), by level of (starting period) household well-being such as poor, non-poor or quintile (indexed by asset index or consumption expenditures per capita), by level of change in availability of water[24]. Conduct and present significance tests.
c. Decomposing changes in outcomes into those due to changes in water availability, those due to changes in sources of water, and those due to quality differences.
d. Create maps of areas of high change in availability, high impact and areas where outcomes continue to lag. Conduct analysis of spatial clustering of impact and hot spots.
e. Discussion of findings.

2) <u>First-difference—regression analysis</u>

Begin with equation 5, modified to reflect time period (t) and include time-invariant unobserved individual effects (ω_i), which might be correlated with the treatment and the error, the dose-response relationship (written in linear form) can be expressed:

$$(8)\quad Y_{it} = \beta X_{it} + \alpha W_{it} + \omega_i + \varepsilon_{it}$$

With the before-after panel survey, we take first differences of the right- and left-hand sides of equation 8.

$$(9)\ Y_{it} - Y_{it-1} = \beta(X_{it} - X_{it-1}) + \alpha(W_{it} - W_{it-1}) + (\omega_i - \omega_i) + (\varepsilon_{it} - \varepsilon_{it-1})$$

Or

$$(9')\quad \Delta Y_i = \beta \Delta X_i + \alpha \Delta W_i + \Delta \varepsilon_i$$

The first difference estimate[25] eliminates the problem of unobserved, time-invariant individual effects (a problematic source of endogeneity). The estimates of α from (9') can be considered an unbiased estimate of the mean impact of the water investment under the assumption that the model is correctly specified and the error term is uncorrelated with the other variables in the equation ($\text{cov}(\varepsilon_{it}, W_{it}) = 0$, $\text{cov}(\varepsilon_{it}, t) = 0$, and $\text{cov}(\varepsilon_{it}, W_{it} * t) = 0$). Unfortunately, these assumptions rarely hold. The last assumption is known as the parallel trend assumption and means that unobserved characteristics affecting water uptake do not vary over time with water availability. As noted above, we have many reasons to suspect this assumption.

We will conduct the analysis as shown in (9') and present the results; this will provide the first quantitative estimate of water investment impacts. However, the assumption of exogenous treatment is limiting—water uptake is likely to be correlated with time-varying unobserved factors—and this correlation is likely to lead to biases. However, for some outcomes (such as time saved hauling water), this bias is likely to be relatively small. We will compare the results from this analysis with those from the summary statistic assessment to highlight commonalities and sources of potential bias.

3) <u>Conduct IV analysis</u>

[24] Please note that our quantitative analysis can disaggregate by any group of interest, although we are not generally able to disaggregate household-level measures by individual characteristics. For example we can compute changes in access (quantity and quality) of water for households, and these can be broken down by household attributes such as poverty status, headship, etc. We could also apportion household-level changes to individuals and compute changes in access to water by, say, gender, but, since we do not observe intra-household distributions of household-level variables, this last apportionment will be approximate.
[25] It might be necessary to correct for serial correlation.

We begin with the specification (9'), which, without being specific about the functional relationship (and suppressing the i subscripts), leads to the following:

$$(10)\Delta Y = \emptyset(\Delta X, \Delta W, \Delta A)$$

Where ΔW is a continuous variable (e.g. quantity of water consumed) that is chosen by the household in a prior economic decision, called the first-stage decision (FS), and ΔA represents changes in unobserved factors. This decision involves observed exogenous factors, or instruments (ΔZ) and unobserved factors, e.g.

$$(11)\Delta W = \mu(\Delta X, \Delta \mathrm{Z}, \Delta V)$$

where ΔV represents changes in unobserved factors affecting the FS decision. Obviously, ΔZ would include the exogenous change in water availability as discussed above, and the FS decision does not depend on the outcome, ΔY. The effect of interest is the impact of ΔWon the outcome.

We assume that the exclusion restriction holds and that the outcome depends on ΔZ only through its impact on ΔW. Under these assumptions, IV estimation of (8') leads to consistent estimates of the treatment effect. As shown by Cameron and Trivedi (p. 884), the LATE estimate (Imbens and Angrist, 1994) can be directly computed from this estimate, in the case where ΔW is binary. The generalization presented by Hoderlein and Sasaki allows us to compute the LASD using similar techniques. This estimate can be used to compute the total impact of the investment, even in the presence of heterogeneous effects.

Thus, the second quantitative estimate of impacts of the water investment will be the IV estimates applied to the first-differenced data. It will be necessary to introduce sufficient flexibility in estimation of 9', as we plan to estimate a number of outcome equations, each of which may have a different functional form. For example, change in diarrhea incidence will be estimated using a health production function (e.g. Rosenzweig and Schultz 1983). See the discussion of the analysis of the first round data (above) for more details.

4) Conduct GPSM analysis

Implementation of the GPS method consists of three steps (Bia and Mattei, 2008). In the first step, we estimate the score r(t, x). In the second step, we estimate the conditional expectation of the outcome as a function of two scalar variables, the treatment level T and the GPS R: β(t, r) = E (Y | T = t,R = r). In the third step, we estimate the dose-response function, μ(t) = E[β{t, r(t,X)}], t ∈ T, by averaging the estimated conditional expectation, $\hat{\beta}${t, r(t,X)}, over the GPS at each level of the treatment.

The Average Dose-Response Function (ADRF) is then obtained by estimating E [Y |T_i , R_i] for every value of t (which implies re-estimating *r* in each stage). To test for the effect being zero, we conduct a joint significance test of the estimated α variables in (4). Under the alternative hypothesis (that at least one of the coefficients is not zero), the F statistic has a non-central F distribution. A non-central F distribution is the ratio of a non-central chi-squared and a (conventional) chi squared random variable.

The continuous treatment approach embodies the central idea that there is only one "group", because all households are treated (there is no control group), and what varies is the intensity of the treatment assigned to each household. Hypothesis testing is complicated by the need to test joint hypotheses (rather than using t statistics, we need F statistics). Software for estimating these effects is directly available in Stata (Bai and Mattei, 2007). By applying the technique to first-differenced data, we eliminate biases associated with unobserved time-invariant household effects.

5) Conduct qualitative analysis

Qualitative data will be collected and analyzed to provide additional insights into the processes and subjective factors which are expected to influence the outcomes and impacts of improved water supplies and which so not lend themselves to quantitative methods. Qualitative data will be generated through semi-structured interviews with key informants, focus group discussions (FGDs) with important target groups, direct observation and case studies in Dar es Salaam and Morogoro. The follow-up qualitative data collection will be implemented approximately one year after the baseline qualitative data collection. As mentioned above, we intend to implement the qualitative research components with the same key informants in both rounds and, where possible, FGD members. The case studies will be presented in narrative form, including relevant background information, photographs and relevant documentary information.

Key Products

As described in the SOW provided to SI by the MCC, the impact evaluation will result in several valuable deliverables. These key products will include a baseline data report, ERR calculations, compact closeout reports, and a final evaluation report and dataset.

Once the first data collection round is complete, a baseline data report will be created that explores how well suited the data are to address the evaluation questions. The report will also highlight data on key outcomes of interest, assess statistical power, compare treatment/control groups, and discuss early lessons learned. It will also contain statistical analysis of outcomes and covariates by gender, by household well-being, and by water availability. This baseline report will also include preliminary qualitative research findings. We will also include maps of water availability. As part of the baseline report, SI will review and update the ERR in comparison to the MCC's baseline ERR.

As the MCA-T compact closes in September 2013, the initial baseline data report and ERR calculations will be expanded to serve as compact closure documentation (requested by June 2013). The updated report will discuss unintended results of the intervention, new lessons learned, questions of sustainability, and other items as requested by the MCC.

Once data collection is complete, SI will generate a final report, dataset, and ERR calculation. These materials will be shared with MCC and key stakeholders for review and comment before drafts are finalized. SI will present and share documents with MCC, MCA-T, and other stakeholders as requested.

Dissemination Plan

SI will adapt the baseline report and initial ERR re-calculations to serve as a Compact Closeout Report in June 2013. We anticipate that both the first-round survey and the initial round of qualitative data collection will be completed in advance of June 2013, providing ample information for a clear discussion of the initial data. We also anticipate that the qualitative research will provide a number of useful narratives that will be compiled and included for dissemination with the Compact Closeout Report. Also, SI plans to conduct 3 to 4 outreach sessions in Tanzania and Washington. The sessions will discuss implementation, lessons learned, and results.

Evaluation Team

The SI evaluation team has several key personnel that will work together to design and implement the IE, analyze the data, and produce final reports. Team composition is detailed in Table 8, as follows:

Table 8: Evaluation Team Positions and Responsibilities

Name	Position	Responsibility
Mike Duthie	Program Manager	Lead engagements with MCC, MCA and partners. Advise IE design methodology and data collection processes. Supervise completion of all deliverables.
Jeffrey Alwang	Principal Investigator	Lead development of IE methodology data collection methodology. Supervise preparation and implementation of each survey data collection wave. Co-lead author of all deliverables.
Charles Pendley	Principal Investigator	Lead development of qualitative data collection plans and methodology. Inform and advise qualitative data collection. Co-lead author of all deliverables.
James Habyarimana	Senior Analyst	Provide comments and revisions for IE design methodology, baseline data report, and final report.
Ralph Hall	Senior Analyst	Revise and make substantial comments to IE design. Advise data collection instrument preparation. Make comments and revisions for all other deliverables.
Eric Vance	Statistician	Along with Jeff Alwang, lead analysis of each data collection wave. Major author of baseline and endline data reports. Contribute heavily to methodology and data sections of final report. Lead ERR updates.
Mark Seiss	Junior Analyst/Statistics	Along with Eric Vance, assist PIs with analysis of each data collection wave. Make heavy contributions to data reports. Assist with ERR updates.

Work Plan

The work plan for the evaluation is outlined in the Table 9 below. The plan accounts for each major deliverable along the expected timeline of the evaluation. Several important elements of the plan, including timing of the endline survey, are dependent upon the construction schedule.

Table 9: Work Plan

Task Area 1: Evaluation Design, Planning, and Peer Review			
Deliverable	**Comments**	**Period Active**	**Due Date**
1. Scoping trip SOW	SI staff will draft a Scope-of-Work for the initial scoping trip and submit for approval to MCC.	January 2012	February 1, 2012
2. Scoping trip report	SI will provide a template for the trip report for approval by the MCC. The PIs and SI staff will report on trip, focusing especially on recommendations for IE design.	March 2012	April 1, 2012
3. Preliminary IE design	Using trip report, SI will develop a preliminary IE design (template or outline to be provided by MCC) which outlines study methodology, data collection plans, and analysis options. SI will submit the report for possible internal peer review by MCC.	March-April 2012	April 15, 2012
4. Final IE design	After receiving comments from MCC, SI will finalize IE design.	May-June 2012	July 1, 2012
5. IE Executive Summary/Website Description	IE design will be summarized and disseminated in executive summary form for non-technical experts to read and understand. This summary will also be translated into Swahili. IE design will also be explained in a one-page summary for posting on the MCC website.	June-July 2012	July 31, 2012
Task Area 2: Evaluation Implementation and Baseline Data Collection Support			
Deliverable	**Comments**	**Period Active**	**Due Date**
6. Support Data Collection Efforts	SI will coordinate with data collection partner, revise/develop data collection instrument as necessary and set systems in place for successful baseline data collection. SI will work with local partner, including through in country support (as required) to facilitate all baseline data collection efforts	September-April 2013	May 1, 2013
7. Baseline data report/data quality report	SI will work with data collecting partner to produce initial report on baseline data. Analyze baseline data, summarize field work, and make recommendations for using baseline data appropriately. SI will use baseline data findings to generate a report on data quality, including discussion of potential concerns and issues to be aware of as field work commences.	March-April 2013	May 1, 2013
Task Area 3: Baseline Data Analysis and Reports			
Deliverable	**Comments**	**Period Active**	**Due Date**
8. Detailed outline of the Compact	Outlines results of baseline data analysis, including outcomes of interest, statistical power, treatment/control group comparisons,	January – March 2013	March 15, 2013

Baseline Report	and any lessons learned.		
9. Draft Compact Baseline Report and updated ERRs	Include elements above and circulate for comment from stakeholders. Update ERR from existing data. Also provide assessment of ERR model and propose any required changes to the original ERR model if necessary.	March-April 2013	April 15, 2013
10. Presentation and feedback from stakeholders on Draft Compact Baseline Report	Including MCC-Tanzania and its implementers	May 2013	
11. Final Compact Baseline Report (including data)	Incorporate feedback from MCC and other stakeholders into final report	May-June 2013	June 15, 2013
Task Area 4: Support to MCA-T between Data Collection Periods			
Deliverables	**Comments**	**Period Active**	**Due Date**
12. Ongoing MCA-T support	Provide guidance to the MCA staff as needed	May 2013 – May 2014	
13. Qualitative research on interim-project results	If necessary and deemed useful in order to describe interim impact results at Compact closeout	May 203 – May 2014	
Task Area 5: Compact Closeout Data Analysis and Reporting			
Deliverables	**Comments**	**Period Active**	**Due Date**
14. ERR Re-calculation	Use existing data with refined model		June 2013
15. Outline for interim evaluation report	Outlines results of existing data analysis, including outcomes of interest, statistical power, treatment/control group comparisons, and so on. Also outlines unintended results, key lessons learned, applicability, analysis of sustainability, and a summary of the existing evaluation results. Since closeout is so close to the time of the baseline data report, the same report will be adapted to serve for the closeout document.		June 2013
16. Plan for a post compact data collection and analysis	Including budget, schedule, sample size, and so on. Will outline the major steps to take place for the final data collection and analysis.		June 2013
17. Draft Final Closeout Evaluation	Using above elements and circulate for comment.		August 2013
18. Final Closeout Evaluation/Data			September 2013

Files			
Task Area 6: Follow Up Data Collection Support			
Deliverables	**Comments**	**Period Active**	**Due Date**
19. Provide on-going support to data collection efforts	As above	September 2013 – April 2014 (Dependent upon construction schedule)	
20. Short data quality report	Discuss quality of data and suggestions to improve future data collection efforts		May 2014
Task Area 7: Data Analysis and Final Evaluation Reporting			
Deliverables	**Comments**	**Period Active**	**Due Date**
21. Recalculate ERR			June 2014
22. Outline of Final Evaluation Report			June 2014
23. Draft Final Evaluation Report			August 2014
24. Final Evaluation Report and Data			September 2014
Task Area 8: Dissemination of Results			
Deliverables	**Comments**	**Period Active**	**Due Date**
25. Outreach Sessions	Conduct 3 to 4 outreach sessions in Tanzania. Based on final reports; ensure that information goes to most relevant parties.		September-December 2014

References

Alwang, J. and E. Gacitua-Mario (2008), “Poverty and Social Impact in the Agricultural Sector: Lessons from Experience,” Development Policy Review 25(2).

ADB (2010), Evaluation of Urban Services and Water Supply and Sanitation Sector in Viet Nam.

AfDB (2010), Project Completion Reports, Integrated Water Supply and Sanitation Project for Nampula and Niassa Provinces, Mozambique and Nchisi and Mzimba Districts, Malawi.

Angrist, J. (2010). Econometrics Teaching Notes. MIT Economics Department Web-site.

Angrist, J., Imbens, G., & Rubin, D. (1996). Identification of causal effects using instrumental variables. Journal of the American Statistical Association , 91.

Bardasi, E., K. Beegle, A. Dillon, & P. Serneels (2010). Do Labor Statistics Depend on How and to Whom the Questions Are Asked? Results from a Survey Experiment in Tanzania, Policy Research Working Paper 5192. Washington, Dc: World Bank, Development Research Group.

Bia, M. & A. Mattei (2008) A Stata package for the estimation of the dose–response function through adjustment for the generalized propensity score, The Stata Journal, 8, Number 3, pp. 354–373.

Cameron, A C. & P. Trivedi (2005). Microeconometrics: Methods and Applications. Cambridge and New York: Cambridge University Press.

Deaton, A. 2010. Instruments, randomization and learning about development. Journal of Economic Literature. v. XLVIII (2):

Desai, S. & Alva, S. (1998), `Maternal education and child health: Is there a strong causal relationship?', Demography 35(1), 71{81.

Doering, Ernst (2006), Reform of the Water Sector in Tanzania, GTZ - Support to the Tanzanian Water Sector Reform Programme.

Galiani, S., P. Gertler & E. Schargrodsky (2005). “Water for Life: The Impact of the Privatization of Water Services on Child Mortality”, Journal of Political Economy, vol. 113, no. 1, 83-120.

Gertler, P.J., S. Martinez, P. Premand, L. B. Rawlings, C. M. Vermeersch (2011). Impact Evaluation in Practice. Washington, D.C.: World Bank.

Hahn, J., Todd, P., Van Der Klaauw, W. (2001), “Identification and estimation of treatment effects with a regression discontinuity design,” Econometrica 69, 201–209.

Heckman, J. J. (1979) “Sample Selection Bias as a Specification Error,” Econometrica, Vol. 47, No. 1, pp. 153–161.

Heckman, J. J. (1990) “Varieties of Selection Bias,” American Economic Review, Vol. 80, No. 2, pp. 313–318.

Heckman, J, H Ichimura, J Smith, & P. Todd. 1998. Characterizing selection bias using experimental data. Econometrica. 66: 1017-99.

Heckman, J. J. & E. Vytlacil (1998) “Instrumental Variables Methods for the Correlated Random Coefficient Model: Estimating the Average Rate of Return to Schooling When the Return is Correlated with Schooling,” The Journal of Human Resources, Vol. 33, No. 4, pp. 974–987.

Heckman, J. J. & E. Vytlacil (2005) “Structural Equations, Treatment Effects, and Econometric Policy Evaluation,” Econometrica, Vol. 73, No. 3, pp. 669-738.

Heckman, J. J. & E. Vytlacil (2007) “Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments” in J.J. Heckman and E.E. Leamer (ed.) Handbook of Econometrics, Vol. 6, Ch. 71.

Hirano, K., & G. W. Imbens. (2004). The propensity score with continuous treatments. In Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives, ed. A. Gelman and X.-L. Meng, 73–84. West Sussex, England: Wiley InterScience.

Heinrich, C, A. Maffioli & G. Vázquez, (2010). "A Primer for Applying Propensity-Score Matching," SPD Working Papers 1005, Inter-American Development Bank, Office of Strategic Planning and Development Effectiveness (SPD).

Hoderlein, S. & Y. Sasaki (2011) Continuous Treatments. Working Paper, Boston College/ Brown University.

Imbens, G. W (2000). “The role of the propensity score in estimating dose-response functions”, Biometrika 87(3), 706-710.

Imbens, G. W. & J. D. Angrist (1994). “Identification and Estimation of Local Average Treatment Effects,” Econometrica, Vol. 62, No. 2, pp. 467-475.

Imbens, G. W. (2000), `The role of the propensity score in estimating dose-response functions', Biometrika 87(3), 706{710.

Iyer, P., Evans, B., Cardosi, J. & Hicks, N. (2005), Rural water supply, sanitation and budget support: Guidelines for task teams, Operational guidance for World Bank group staff, The World Bank.

Jalan, J. & M. Ravallion (2003). "Does piped water reduce diarrhea for children in rural India?," Journal of Econometrics, vol. 112(1), pages 153-173.

Kjellén, Marianne, (2009), Structural Leakage in Dar es Salaam: The Investment Deficit in Water Distribution, Stockholm Environment Institute (SEI).

Lechner, M. (2002), `Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies', The Review of Economics and Statistics 84(2), 205-220

Lee, L., Rosenzweig, M. R. & Pitt, M. M. (1997), `The effects of improved nutrition, sanitation, and water quality on child health in high-mortality populations', Journal of Econometrics 77(1), 209-235.

Merchant, A. T., Jones, C., Kiure, A., Kupka, R., Fitzmaurice, G., Herrera, M. G. & Fawzi, W. W. (2003), `Water and sanitation associated with improved child growth', European Journal of Clinical Nutrition 57(12), 1562-1568.

MCA-Tanzania (2009), Gender Policy.

MCA-Tanzania (2010), Monitoring and Evaluation Plan.

MCA-Tanzania (2012), Monitoring and Evaluation Plan.

Napacho, Z. A. and Manyele, S.V. (2010), Department of Chemical and Process Engineering, College of Engineering and Technology, University of Dar es Salaam), Quality assessment of drinking water in Temeke District (part II): Characterization of chemical parameters.

Nauges, C., D. Whittington (2010). "Estimation of Water Demand in Developing Countries: An Overview," World Bank Research Observer, vol. 25, no. 2, 263-94

Novak, Lindsey K. (2011), The Impact of Access to Water on Child Health in Senegal, The International Initiative for Impact Evaluation DRAFT.

Pitt, M & S.R. Khandker (1998). "The impact of group-based credit programs on poor households in Bangladesh: does the gender of participants matter?" Journal of Political Economy. 106(5): 958-996.

Pigeon, Martin (2012), From Fiasco to DAWASCO, Remunicipalising Water Systems in Dar es Salaam, Tanzania.

Rosenzweig, M. R. & T. P. Schultz (1983) “Estimating a Household Production Function:
Heterogeneity, the Demand for Health Inputs, and Their Effects on Birth Weight,”Journal of Political Economy, Vol. 91, No. 5, pp. 723–746.

Rosenbaum, P., & D. Rubin. 1983. “The Central Role of the Propensity Score in Observational Studies for Causal Effects”. Biometrika 70(1): 41-55.

Rubin, D. 1974. "Estimating Causal Effects to Treatments in Randomized and Nonrandomized Studies". Journal of Educational Psychology. 66: 688-701.

Rubin, D. 1977. "Assignment to Treatment Group on the Basis of a Covariate". Journal of Educational Studies. 2:1-26.

Rubin, D. 1979. "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies". Journal of the American Statistical Association. 74: 318-328.

Rubin, D., & N. Thomas (2000), "Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates," Journal of the American Statistical Association 95: 573-585.

Stuart, E. A. and D. Rubin, 2007 “Matching methods for causal inference: Designing observational studies” To appear in Best Practices in Quantitative Methods, J. Osborne (Ed.). Thousand Oaks, CA: Sage Publishing.

UN-Habitat (2007) Water for African Cities (WAC) Dar es Salaam, Evaluation Report.
UNWater (2008), Tackling a global crisis: International year of sanitation 2008, Full report, The United Nations Department of Economic and Social Affairs

UNICEF. 2002. Household water security e-conference and virtual forum.” Internet, available at http://www.unicef.org/french/wash/files/HHrep.pdf. Accessed 5 May 2012.

Waddington, H., Snilstveit, B., White, H. & Fewtrell, L. (2009), Water, sanitation and hygiene Interventions to combat childhood diarrhoea in developing countries, Systematic review, International Initiative for Impact Evaluation (3ie).

Water Aid (2002), Using the Household Budget Survey, the Demographic and Health Survey and population censuses to measure trends in water and sanitation, Ministry of Water and Livestock Development, in collaboration with the Eastern Africa Statistical Training Center.

Water and Sanitation Program (WSP) Tanzania (2011), Tanzania: A Handwashing Behavior Change Journey.

Water and Sanitation Program (WSP) Tanzania (2011), Water Supply and Sanitation in Tanzania, Turning Finance into Services for 2015 and Beyond, AMCOW Country Status Overview.

Wegelin, W A, McKenzie, R.S., Van Der Merve, B., Maboyja, M. (2009), The Emfuleni Water Loss Project - A Major Challenge, Proceedings of the 5th IWA Water Loss Reduction Specialist Conference, International Water Association (IWA), Pages: 409-416.

Whittington, D.; D. T. Lauria, & X. Mu (1991). "A Study of Water Vending and Willingness to Pay for Water in Onitsha, Nigeria", World

WHO/UNICEF (2006), `Core questions on drinking-water and sanitation for household surveys'. Geneva: WHO Press.

WHO (2010), Burden of disease and cost-effectiveness estimates, Technical report, World Health Organization.http://www.who.int/watersanitationhealth/diseases/burden/en/.

World Bank (2011), Implementation Completion and Results Report, Dar es Salaam Water Supply and Sanitation Project.

Annex I: Project Logic

WATER SECTOR PROJECT LOGIC

Lower Ruvu Plant Expansion and Morogoro Water Supply

PROCESS

Activities	Indicators
Finance feasibility, design activities	**Value of feasibility/design contract ($)*** **Value of feasibility/design contract disbursed ($)*** Certificate for Environmental Impact Assessment (EIA) issued (Date)
Finance construction activities	**Value of construction contract ($)*** **Value of construction contract disbursed ($)***

OUTPUTS

Result	Indicators
Improve treatment plants	**Schedule of Performance Ratio (ratio)**
Increase water production	**Volume of water produced (liters/capita/day)***
Reduce water losses	**Non-revenue water (%)***
Improve financial sustainability	**Operating Cost Coverage (ratio)***
Increase temporary employment	**Total number of people temporarily employed/contracted by MCA-IEs (#)***

OUTCOMES

Result	Indicators
Improve water service coverage	**Number of non-domestic customers (#)*** **Number of domestic customers (#)*** **Percentage of non-active customers to total customers (%)** Percentage of households with access to improved water supply (%)
Improve quality of service	**Continuity of service (hours/day)***
Improve quality of water	**Nephelometric turbidity units (NTU)** Coliform Microbial Density (per 100 milliliters) Free Chlorine Residual (FRC)
Increase water consumption	**Volume of commercial water consumption (cubic meters per month)*** **Volume of residential water consumption (liters/capita/day)***

SHORT-TERM OBJECTIVES

Result	Indicators
Decrease incidence of water-borne related morbidity	Percentage of population with diarrhea in the last 2 weeks (%)
Improve human capital accumulation	Average hours worked last week (hours) Percentage of school children who missed any in the last 4 weeks (%) Average time spent fetching water from home in last week (min)

MEDIUM-TERM OBJECTIVES

Result	Indicators
Decrease in mortality	National level <5 mortality rate (per 1000 births) National level Adult mortality rate (per 1000)
Increase investment and economic activities	Average current value of household assets per capita ($) Average value of commercial assets ($)

COMPACT GOAL

Result	Indicators
Poverty Reduction and Economic Growth	Average annual household income per capita ($)

Bolded text refers to Indicator Tracking Table (ITT) Indicators which will be reported on a *quarterly* basis. All other indicators will be reported on as data is available.

* Refers to Millennium Challenge Corporation Common Indicators for the Water Sector

Annex II: NPS Analysis

MCA-T-sponsored oversampling in Dar es Salaam and Morogoro during 2010 Tanzania National Panel Survey
An evaluation of its usefulness for impact evaluation

Background

The Tanzania National Panel Survey (NPS) is a nationally representative and multi-year initiative led by the Tanzania National Bureau of Statistics (NBS). The survey is designed to provide high-quality household-level data to monitor and understand poverty dynamics and improve the understanding agriculture and linkages between farm and non-farm activities. The NPS is an integrated survey, covering a range of socioeconomic factors, and its breadth allows it to be used for multiple purposes. Since it is a panel survey, the NPS serves as a useful tool for analysis of changes over time. While originally intended to aid in the study of poverty and welfare dynamics, it is readily adoptable to the analysis of other changes. One such change would be a change in access to water or other public services.

Each wave (round) of the survey is fielded over a period of 12 months, with 1/12 of the sample being enumerated in each month. The first round was conducted between October 2008 and September 2009. The second round occurred during October 2010 through September 2011. The national sample size is 3,265 households in 409 enumeration areas. The sample includes 555 households within 70 clusters from Dar es Salaam and 112 households in 14 clusters in Morogoro. The sample is representative at the urban/rural level and at the level of the major agro-climatic zones.

The first round survey instruments included Household, Agricultural and Community Questionnaires. Information is collected on multiple levels: individual-level information (such as labor, education, health), household-level information (such as water and sanitation, food consumption expenditures, asset ownership, housing amenities), and farm-level information for plots, crops, and other activities.

The Household Questionnaire is detailed and multi-topic. In the 2010/11 survey, this questionnaire included modules on demographics, education, health, labor, subjective welfare, governance, consumption, housing/water/sanitation, expenditure, assets, social safety nets, credit, welfare shocks, deaths of household members and anthropometric measurements. The Community Questionnaire collects information on physical and economic infrastructure and community-wide trends.

The latest information from NBS is that the full second round data will be available in June 2012.

The section on Housing, Water and Sanitation (section J) within the Household Questionnaire was expanded during the 2010/11 round with additional funding from MCA-T to include questions on water source, availability, water quality and information about the local water authority. The sample was also expanded in an attempt to produce a baseline for the MCA-Ts water supply investments for Dar es Salaam and Morogoro.

MCA-T's NPS Sample

Between August 25, 2010 and October 3, 2010, MCA–T contracted NBS to collect survey data in 80 clusters in Dar es Salaam, Morogoro and Pwani. These data are referred to as the MCA-T data. The MCA-T data were expected to serve as a baseline for MCC's Water Sector Impact Evaluation in Dar es Salaam and Morogoro. Enumeration of the MCA-T data was part of training for the second round NPS, and the survey instruments were the same as those used in the second round of the NPS. The 80 MCA-T clusters are, thus, in addition to those forming a regular part of the NPS. Like the core NPS sample, it relies on two-stage cluster sampling with 8 households per cluster.

The *Social Impact* team conducted an evaluation of the MCA-T data and concludes that the data are not ideally suited for use as a baseline in a robust evaluation of Water Sector Investment impacts. This conclusion is based on three factors: (i) the timing of the survey; (ii) its coverage, including sample size and geographical focus; and (iii) the questions and information available from the survey.

The *Social Impact* evaluation involved a comprehensive review of supporting documentation, discussions during a scoping mission with NBS and other entities involved in data collection, discussions with the water authorities in Dar es Salaam and Morogoro, and an evaluation/analysis of the raw data.

Timing of the MCA-T data collection

Data collection for the MCA-T oversample occurred during August and September 2010. This timing presents two problems. First, since in Dar es Salaam the Lower Ruvu water project is not expected to begin operations before April/May 2013, the gap between data collection and start of a plausible impact will be nearly three years in Dar es Salaam. Several intervening events have affected water supply in the area, including new borehole sources, improved distribution infrastructure, and additional connections. The delay between 2010 and 2013 is likely to compound attribution problems as multiple changes to the system have occurred; in principle, these changes can be documented, but their impacts will be difficult to separate from the impacts of the MCA-T investments. A general principle of impact assessment is to collect baseline data as close to the time of impact as possible. Morogoro impacts are likely to be felt earlier (as early as June/July 2012), so the gap issue is not as critical for that intervention.

The second problem has to do with the short time window associated with the oversample. As noted, the NPS was designed to be representative at sub-geographies (e.g. urban/rural) by collecting information over an entire year from each of these sub-geographies. This collection method helps eliminate the serious problem of seasonality. By using a sample design and survey instruments designed to span a year and compacting data collection into a two-month period, intra-year representativeness of the sample is compromised. This is a special weakness in the case of water-related changes whose impacts exhibit pronounced seasonality corresponding to the March-May (long) and Nov-Dec (short) rainy seasons. Two week recall of key water-related diseases among children and month-long recall of water expenditures (discussed in more detail below) will not allow the instrument to reflect this seasonality. This is a critical timing issue: the baseline NPS data were not collected during periods when water-related diseases are most common and an evaluation relying on these data will not be able to detect an important expected impact of the investment.

Sampling issues

As noted in the RFP for the impact evaluation (citation), the methodology implemented to select MCA-T clusters and sample households within them is unclear. The sampling framework ideally incorporates information on the specific intervention and where potential impacts will/will not be observed. The map presented with the RFP indicate areas within the city where current water availability is deemed regular (24-hour availability), others experiencing different rationing levels and others with no availability (p. 26 of RFA). This map was created following discussion with DAWASA personnel, yet none of the parties whom we interviewed acknowledged responsibility for it. As a map of service reliability, it is not accurate. In addition to this problem, the sample frame should be based on an assessment of zones of influence of the change in water supply, not a static estimate of current supply conditions. Thus, the underlying geography of the oversample is suspect. We conclude that the procedure implemented for selection and sampling for the MCA-T data will not be consistent with the research study, making it difficult to classify observations as treatment or control.

A second concern relates to the sample size. The maximum number of households surveyed (not addressing the problem of missing variables) in the oversample is 296 for Dar es Salaam and 320 for Morogoro. While in principle, these could be added to the remainder of the 2010/11 NPS sample, intra-year representativeness will still be lost, and the resulting sample size is still relatively small. The sample size issue is likely to be extremely critical as numerous confounding factors may make tenuous the link between increased water supply and measurable outcomes. The power calculations presented in the RFA are not convincing and the small sample is likely to make it impossible to detect changes in water-related diseases and illness, school attendance, expenditures on waters, and other outcomes.

On their face, the power calculations for minimum sample size were not done correctly, mainly because the two outcome variables used were not appropriate. A power calculation usually begins with identification of the most important outcome indicators that the program is designed to improve. The calculations in Annex V of the RFA used household per capita consumption and access to piped water. The MCA-T is not investing in piping to households, so any change in this variable would not be a direct result of the intervention, and consumption expenditures may actually fall (as households spend less on alternative sources of water—likely other expenditures would substitute for the decline in water expenditures). Total consumption expenditures are usually taken as a measure of household well-being, and the use of this variable to conduct the power calculation might be based on a notion of improvements in household well-being over time from better water availability. However, this variable is only likely to change relatively slowly as increased time availability and improvements in human capital lead to more incomes and eventually raise well-being. Neither of these “outcomes” is an appropriate short-run indicator of MCA-T investment outcomes. Instead, rates of diarrhea, school attendance, time spent gathering water and, possibly, expenditures on water should have been used to conduct the power calculation.

Another problem is that the concept of "treatment" and "control" groups used in the RFA is not relevant for the evaluation: the treatment (enhanced availability of water) is a continuous variable and everyone in the interconnected system will be affected. The evaluation will need to take a continuous treatment approach, which requires increased sample size. The RFA suggests contacting the water authorities in Dar es Salaam and Morogoro to identify treatment and control areas and then overlay this information with the MCA-T GPS data. The Social Impact team has done this, and concludes that the MCA-T data are inadequate.

Questionnaire content and survey information
A final concern about the MCA-T oversample relates to the adequacy of the questions in the questionnaire. Section J of the questionnaire addresses housing, water and sanitation. The questions on water access contain a good detail on household sources of water, distance (in time) to the sources, connection to the water grid (including billing information for a small proportion of the sample), satisfaction with public piped water, and access and use of water from sources outside of the household (e.g. neighbors, kiosks, etc.). The information from the questions provides a good portrait of sources and their diversity, but some important information is missing. Missing information includes:

1) Information on who is responsible for hauling water. As time savings is likely to be an important intermediate outcome from the MCA-T investment, it is important to know whose time is being saved. This information will help us value that person's time and create a stronger link between time savings and other outcomes such as school attendance, productive labor, etc.

2) Information on the volume of water consumed by the household. For piped water, this information is available for only 7 % of Dar households (who were able to show a bill). For sources outside the household (i.e. not piped into the home), we have information on what type of container is used to retrieve the water, frequency of fetching water, and the price from various sources (per 20 liters). However, because we do not know the volume of the water in the container, we cannot reconstruct the total volume consumed nor the total amount expended on water. The expenditure component of the survey (section L) asks about total amount spent on water in the past month. This information might be used in conjunction with the cost per 20 liter information in section J to create a measure of volume of water consumed, but, since most households purchase from numerous sources, the estimate would be at best rough. In addition, because the water question in section L is just one item in a long list of expenditure items, there is likely to be important recall bias in this critical variable.

An obvious advantage of the NPS is the breadth of topics covered. This breadth is especially useful when considering potential impacts of the MCA-T investments. Impacts are likely to be manifest in health outcomes, changes in labor force behavior and educational participation. The survey questions in the health section (section D) are adequate, but outcomes like childhood diarrhea are likely to exhibit strong seasonal patterns. Since the survey was not conducted during the rainy season, variables reflecting a 2 week recall of diarrheal and other water-related health episodes are less likely to fall due to the intervention, compared to the same variable measured during the rainy season.

Annex III: Document Review

Background

To more adequately explore the Tanzanian context, SI has extensively reviewed relevant data and other documentation. Sources include, but are not limited to, the following:

- MCC and MCA-T documents and data, including documents prepared by consultants and contractors engaged by MCC and MCA-T. The most important of these are the NPS and the subsequent oversample of the NPS carried out by NBS on behalf of MCA-T.
- Reports and data from official sources; e.g. MOW, DAWASA, DASWASCO, Ministry of Health and NBS.
- Reports from other donor-supported programs and projects in the urban water sub-sector in Tanzania, e.g. Water Sector Partnership Program (EU), Water sector Development Program (World Bank), GIZ, etc.
- Reports from relevant organizations, research institutions, NGOs, and others, e.g. Ifakara Health Institute, National Institute for Medical Research (NIMR), University of Dar es Salaam, Water Aid, CARE, Plan International, among others.

Details of key documentation obtained and reviewed by SI are presented in the reference section at the end of this report. The literature on urban water supply tended to fall into categories that focused on infrastructure, institutions, interventions and/or individual choice. Sub-genres included Willingness-to Pay (WTP) studies, institutional analysis, water sector assessments, program and project documents, social and demographic surveys, small area and qualitative studies, among others.

While many of the documents reviewed were primarily descriptive in nature, a few attempted to establish associations and correlations between water-related behavior and structural or other external influences. Altogether, more than 150 documents were obtained and reviewed by the team.

Literature on infrastructure tends to focus on the condition and functioning of the physical components of the water supply system/network, which was in most cases found to be deficient in a number of areas, including the treatment facilities, pipe network, meter maintenance and individual service connections.

The literature on institutions includes assessments of the effectiveness and efficiency of the relevant water supply agencies, while the literature that focuses primarily on interventions consists mainly of program and project documents and consultant’s reports. Documents and reports that focus on individual choice presents the perspectives, behaviors and coping strategies of present or future beneficiaries or customers, including low-income groups. Studies employing a qualitative methodology tend to be found in this category.

Infrastructure

Project completion reports from the World Bank (2010) and ADB (2010) cite the poor and neglected condition of the physical water supply infrastructure as a major justification for providing technical and financial support to upgrade and extend the coverage of present urban water supply systems. Project completion reports tend to indicate that physical upgrading of system components was assessed to be among the more successful components of urban water supply projects in terms of completion.

Institutions
A number of project completion/evaluation reports focus on the importance of the efficiency and effectiveness of sector institutions, (World Bank (2011), ADB (2010), Kjellén (2009), UN-Habitat (2007), and Doering (2006) for achieving success, sustainability and the desired impact of urban water supply projects. Even if physical constriction and other targets are achieved, it was observed that benefits and impact may not be achieved or sustained due to managerial weaknesses and human and technical capacity constraints in the responsible sector agencies. Unless sector institutions were strengthened alongside technical improvements, sustainability and lasting benefits would not be achieved.

Interventions
The results of technical and institutional analyses were used as the rationale for designing interventions that were in most cases supported by loans or grants from external funding agencies. Typical project designs included a combination of technical and institutional interventions which were intended to be mutually supportive. However, it was often found that during implementation technical interventions tended to proceed independently of institutional interventions (World Bank, 2011), which were either not implemented or scaled back significantly.

Individuals
Some of the documents reviewed (WSP, 2011, Kjellén, 2009), focus on the role individual/customer's needs and choice play in determining water accessibility and use and customers' responses to urban water supply projects. Also in this category are documents (UN-Habitat 2007) advocating greater stakeholder participation in project decisions to create a sense of ownership and commitment to use and maintain facilities properly.
What was generally found to be lacking in the literature reviewed is a rigorous analysis using a known and randomly selected sample which establishes a link between infrastructural, institutional, intervention and individual parameters of urban water supply projects, which in the SI's view are closely interrelated and will be treated as such in the present IE.

Health Literature
There is a relative paucity of epidemiological studies on water-related diseases in the planned impact area. While diarrheal and other water-related diseases received a significant amount of attention internationally in the 1980's and 1990's, more recently such pandemic diseases as HIV/AIDS, malaria, and tuberculosis have largely eclipsed diarrheal diseases as a subject of research interest. This trend has been reinforced by large amounts of funding from governments, international aid agencies and private foundations. A paper (Napacho and Manyele 2010) presents a study on drinking water quality in Temeke District (Dar es Salaam), which involved analyses of chemical parameters of drinking water samples from different drinking water sources. The drinking water sources examined included tap water, river water and well water (deep and shallow wells). Water quality studied includes pH, chloride, nitrate and total hardness.

The concentration of total hardness in mg $CaCO_3$/L and chloride were obtained by titration method while nitrate concentration levels were determined by spectrophotometer. Tap water was found to be of higher quality than other sources in terms of chemical characteristics. The study revealed that the chemical parameters of water sources did not meet WHO and Tanzania Bureau of Standards (TBS) standards. It was revealed that most of the samples contained chloride levels above allowable WHO limits. It was recommended that drinking water sources for domestic use should be protected from potential sources of pollution**.**

Literature Gaps

In spite of the large number of studies and reports and the amount of available information and data concerning urban water supply in Tanzania, gaps and other shortcomings still exist. Evidence gaps of relevance to the present IE include:

- Challenges in identifying and isolating probable impact areas of the planned water supply improvements in both Dar es Salaam and Morogoro
- Lack of reliable and accurate time-series data on water production and availability, particularly in Morogoro
- Reliable data on incidence and prevalence of water-related diseases in the probable impact areas with specific links to improved water supply
- Lack of adequate data on the definition, incidence and distribution of poverty in the probable impact areas which link poverty to water-related indicators
- Lack of gender disaggregated data on water-related behavior
- Lack of linkage between and compatibility of the available datasets, making establishing associations and correlations of improvements in water supply with other relevant parameters difficult, if not impossible.

Annex IV: Sample Size Calculations

Introduction

The SI team conducted an analysis of calculations for the minimum sample required to produce reliable statistical estimates for MCA-Tanzania's water sector program impact evaluation. We assume a clustered, quasi-randomized evaluation design with a continuous treatment whose magnitude varies by cluster. We also assume that data collection will occur before and after inauguration of the projects. Because of the differences in timing of project completion, we are treating Dar es Salaam and Morogoro as distinct entities in our impact assessment. As a result, we are computing separate sample sizes for each.

The purpose of the sample size estimates is to determine the minimum impact, Δ, that can be detected for a given number of clusters in the sample, *g*, and households in each cluster, *m*, for the evaluation sample. If the impact of the treatment is at least as large as Δ, we will be able to detect it 80 percent of the time in a sample of total size *mg*. If the treatment impact is less than Δ, we are less likely to detect it, although it is still possible. Initially, we assume *m*=8, following the NBS convention of 8 observations per cluster in the HBS and NPS samples. We also allow *m* to vary to examine its effect on sample size.

The optimal sample size depends, among other things, on the focal outcome variable. In determining the sample size, we used several alternatives: household expenditures on water, time spent collecting water (wet season and dry season), household exposure to serious water shortages, and indicators of diarrheal disease among children. These indicators were chosen because they coincide with the project logic as spelled out in the M&E plan and information on them was available from the recent NPS MCA-T household survey.

Discrete treatment

We begin with a simple exercise assuming that the treatment is discrete. As noted above, we plan on using difference-in-difference estimators with repeat observations on individual households. Murray (1998, chapter 9) generated formulae for power calculations for this type of survey, and we follow his procedures. The main analysis is based on the following equation (number 9.23 in Murray's book):

$$g = \frac{2\left(1 + (m-1)\widehat{ICC}\right)(t_{\alpha/2} + t_{\beta})^2}{m\hat{\Delta}^2}\hat{\sigma}_y^2$$

Where:

- g: number of clusters in each condition (treatment/control)
- m: number of observations per cluster
- ICC: Intracluster correlation
- α: type I error rate
- ß: type II error rate
- $\hat{\sigma}_y^2$: estimated variance of the outcome variable

To implement this equation, we need information on these parameters. Using the NPS convention, we begin with m=8; we also examine sensitivity to alternative observations per cluster. We assume Type I and II error rates of 5% and 20% respectively and a change in outcome variables of at least 20%. The above equation ignores the panel nature of our data and the likely impact of inter-period correlation. However, ignoring these potential sources of correlation allows us to generate conservative estimates of necessary sample sizes. We examine the sensitivity to these assumptions by

Intra-cluster correlation

The most controversial issue in sample design is the intra-cluster correlation (ICC), defined as $\rho = \frac{\tau^2}{\tau^2+\sigma^2}$, or the proportion of overall variance explained by within group variance[26]. We use the NPS 2010 MCA-T sample to compute the ICC for the outcome variables described above.

Continuous treatment

The continuous treatment case alters the calculation slightly. Murray (1998) shows that with a sufficiently large sample size t=2.80 will guarantee a power of 80%. To achieve this, we need to ensure that $\hat{\beta} \geq 2.80 s.e.(\hat{\beta})$, and using the definition of var($\hat{\beta}$) and knowledge that $\sigma_\varepsilon^2 = \sigma_{Y|T}^2$and $\sigma_{Y|T}^2 = \sigma_Y^2(1-\rho_{TY}^2)$, we see

$$\left(t_{\alpha/2} + t_\beta\right) = \frac{\hat{\beta}}{\sqrt{\frac{\sigma_Y^2(1-\rho_{TY}^2)}{n\sigma_T^2}}}$$

And, taking into account the intra cluster correlation (as the survey is clustered), we come up with:

$$\left(t_{\alpha/2} + t_\beta\right) = \frac{\hat{\beta}}{\sqrt{\frac{\sigma_Y^2[1+(c-1)ICC](1-\rho_{TY}^2)}{n\sigma_T^2}}}$$

Where T represents the continuous treatment, and c is the number of clusters, where the number of clusters is given by the results of the discrete treatment case with 8 observations per cluster. The total sample size, under these conditions is determined as

$$n = \frac{(1-\rho_{TY}^2)\sigma_Y^2[1+ICC(c-1)]}{\left[\frac{\hat{\beta}}{\left(t_{\alpha/2} + t_\beta\right)}\right]^2 \sigma_T^2}$$

[26] The ICC can be estimated using a linear treatment model: $Y_{ij} = \alpha + BT + v_j + \omega_{ij}$, where j indexes the cluster and i indexes the individual or household, Y is the outcome and T is the treatment. We assume that clusters are of identical sizes and v_j is IID with variance τ^2, and ω_{ij} is also iid, with variance σ^2. See Dulfo, Glennerster and Kremer (2008), particularly 3921-3923.

The assumptions and data used for the calculations are summarized in Table 4.1.

Sample size estimates are shown for different outcome variables and different assumptions about the number of observations per sampling cluster. Outcome variances are higher in Morogoro, necessitating a larger sample size for the same level of precision, and for reasonably sized samples, we will have a difficult time detecting differences for many of the variables[27].

Results

Results from the power analyses are shown in Tables 4.2 and 4.3. These results show the minimum sample size for a clustered survey with various cluster sizes (in the case of the binary treatment—Table 4.2) and for the continuous treatment with cluster size of eight (Table 4.3). In order to obtain data representative of different seasons, we recommend collecting one half of the sample during the rainy season and one-half during the dry season[28]. A sample size of 2500 would provide sufficient power to detect statistically significant changes in three important outcomes: household expenditures on water, time necessary to haul water in the rainy season, and changes in exposure to major water shocks. Differences in time spent hauling water during the dry season (more time is spent in the dry season) would be harder to detect with a sample size of 2500, but we might be able to detect the difference. Differences in rates of children's diarrhea would be extremely difficult to detect; reasonable power to detect such differences would require sample sizes beyond the capability of this evaluation.

[27] Assuming the patterns of variation and covariation hold in the new sample.

[28] The recall responses from the MCA-T NPS oversample show that there is pronounced seasonality in water sources, travel time to the sources, and other factors related to water purchases. This sample was conducted in August-September, a dry-season period in both cities, but asked recall questions about water sources and time to the sources for both rainy and dry seasons. To obtain representativeness of the rainy and dry season, the sample sizes would need to be roughly doubled.

Table 4.1: Parameters used in sample size calculations

	Mean		σ_Y		$\hat{\beta}$		ρ_{TY}		ICC	
Variable	Dar es Salaam	Morogoro	Dar es Salaam	Morogoro	Dar es Salaam	Morogoro	Dar es Salaam	Morogoro	Dar es Salaam	Morogoro
Water expenditure	82302	6067	8230	5952	26.344	13.17	0.127	0.030	0.04	0.04
Time to collect water, rainy season	14.497	10.756	14.497	14.470	0.041	0.116	0.101	0.150	0.075	0.075
Time to collect water, dry season	24.035	23.482	24.035	33.963	0.069	-0.151	0.036	0.007	0.075	0.075
Water shortage major shock in past 5 years	0.486	0.263	0.486	0.441	-0.004	-0.002	0.301	0.214	0.075	0.075
Water shortage most important shock in past 5 years	0.239	0.034	0.240	0.182	-0.001	-0.000	0.171	0.140	0.075	0.075
Under-five diarrhea in past 14 days	0.095	0.069	0.293	0.253	0.000261	0.000261	.016	.129	.05	.05
Average hours per week water availability (T)	50.6	71.6								

Notes: $\hat{\beta}$ comes from a linear regression of the outcome variable on the treatment. ICC values were approximated based on the regression results from footnote 1. In all cases, we chose a conservative version of the ICC.

Table 4.2: Sample size estimates, binary treatment

Outcome Variable	m=8		m=16		m=32	
Dar es Salaam	c=2g	n	c=2g	n	c=2g	N
Water expenditure	74	593	46	741	32	1038
Time to collect water, rainy season	105	840	68	1089	50	1586
Time to collect water, dry season	136	1091	88	1414	64	2060
Water shortage major shock in past 5 years	195	1556	136	2168	106	3393
Water shortage most important shock in past 5 years	1829	14629	1274	20384	997	31896
Under-five diarrhea in past 14 days	1003	8024	650	10400	474	15168
Morogoro						
Water expenditure	95	763	60	953	42	1335
Time to collect water, rainy season	214	1709	149	2381	116	3725
Time to collect water, dry season	247	1975	172	2752	135	4306
Water shortage major shock in past 5 years	333	2660	232	3707	181	5801
Water shortage most important shock in past 5 years	3325	26601	2317	37067	1812	57999
Under-five diarrhea in past 14 days	1419	11352	920	14720	670	21440

Table 4.3: Sample size for continuous treatment

Outcome Variable			
	Dar es Salaam	Morogoro	Total
Water expenditure	569	1182	1751
Time to collect water, rainy season	1146	314	1460
Time to collect water, dry season	2030	1119	3149
Water shortage major shock in past 5 years	522	1776	2297
Water shortage most important shock in past 5 years	11133	16150	27283
Under-five diarrhea in past 14 days	174431	80602	255033

Annex V: Alternative Methods

Option 1: Regression Analysis

One approach to controlling for initial differences between treatment and comparison groups is to specify a regression model, which includes potentially confounding explanatory variables, such as distance to the pipeline, education, and baseline health outcomes, as independent variables. Assuming that the outcome of interest is measured at the individual level[29] (such as number of school/work days missed due to illness), the basic linear regression model that controls for baseline characteristics and household fixed effects can be written as:

$$(1)\quad y_{ih,F} = \beta' X_{ih,B} + \alpha T_{ih} + \eta_h + \varepsilon_{ih}$$

where $y_{ih,F}$ is the outcome of interest (analysis can be conducted separately for each outcome of interest) for individual *i* in household *h* measured at the follow-up survey; $X_{ih,B}$ is a vector of baseline characteristics, including baseline level of the outcome in question as well as distance to the water pipeline; T_{ih} is a variable representing the level of treatment (described above); η_h is a household-specific error term; and ε_{ih} is an individual level random error term. α represents the average impact of the water program. This model can also be modified to estimate the impacts on important sub-groups. Although this method represents an improvement over the standard before-after and difference in difference designs, it has at least three drawbacks:

1. It relies on the assumption that conditional on the control variables (in other words, for individuals with the same values on the control variables), the treatment and comparison groups would have the same expected outcomes (conditional independence). However, regardless of the comprehensiveness of the baseline data collection, it is possible that some unobserved or uncontrolled differences between treatment and comparison remain. This is typically less of a concern when program participation has more to do with implementing agency placement, as is arguably the case for water infrastructure projects[30], than a household participation decision, where individuals choose whether or not to participate in a program. Nevertheless, unobserved selection bias is more likely under regression analysis than the GPSM design discussed below, as control variables in regression analysis are limited to predictors of outcomes, yet Rubin and Thomas[31] suggest that variables with weak outcome predictive value, including some covariates of participation used in GPSM, are still useful for reducing bias in estimating causal effects.
2. Rubin and Thomas also find that regression analysis using full data sets are subject to more bias than analysis on matched sub-samples, so we may be able to reduce bias by excluding from analysis those comparison units who are very dissimilar to the treatment units, as is done in matching and regression discontinuity designs.
3. Regression analysis assumes the form of the relationship (for example, a linear relationship in the model above) between treatment, treatment effects, and control

[29] For household level changes, the household fixed effects could be removed, or replaced with village or community level fixed effects.

[30] Although households may make the decision to move in or out of the treatment area.

[31] Rubin, Donald B., and Neal Thomas (2000), "Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates," Journal of the American Statistical Association 95: 573-585.

variables, yet there may be important interaction effects or non-linear relationships.

The simple regression approach, however, can be improved upon by noting that the treatment (access to water) is related to the endogenous household decision of water consumption (volume and quality), and this household consumption decision is structurally related to other outcomes. A structural model may aide us in estimating the dose-response relationship between the endogenous consumption decision and other outcomes of interest.

Annex VI: Tanzania Water Quality Standards

Table 6.1: Microbiological Requirements

Class of Piped Water/ Type of test count	**Coliform count per 100 ml at 37 ºC**	**E. Coli (fecal coliform) count per 100 ml at 44 ºC**
Excellent	0	0
Satisfactory	1 – 3	0
Suspicious	4 – 10	0
Unsatisfactory	More than 10	1 or more

Table 6.2: Chemical and Physical Limits for Drinking Water Sources

No.	Name of Constituent	Symbol	Units	Limits
	Toxic			
1	Lead	Pb	mg/l	0.01
2	Arsenic	As	mg/l	0.05
3	Selenium	Se	mg/l	0.05
4	Chromium	Cr	mg/l	0.05
5	Cyanide	Cn	mg/l	0.20
6	Cadmium	Cd	mg/l	0.05
7	Barium	Ba	mg/l	1.00
8	Mercury	Hg	mg/l	0.001
9	Silver	Ag	mg/l	Not mentioned
	Affecting Human Health			
1	Fluoride	F	mg/l	1.5 - 4.0
2	Nitrate	NO_3	mg/l	10 – 75
	Organoleptic			
1	Color		mg/l	15 – 50
2	Turbidity		mg/l	5 – 25
3	Taste		-	Not objectionable
4	Odor		-	Not objectionable
	Salinity and Hardness			
5	pH			6.5 – 9.2
6	Total Filterable Residue		mg/l	500 – 2000
7	Total Hardness	$CaCO_3$	mg/l	500 – 600
8	Calcium	Ca	mg/l	75 – 300
9	Magnesium	Mg	mg/l	50 – 100
10	Magnesium + Sodium sulphate	$Mg\text{-}Na_2$	mg/l	500 – 1000

11	Sulphate	SO_4	mg/l	200 – 600
12	Chloride	Cl	mg/l	200 – 800
	Less-toxic Metals			
13	Iron	Fe	mg/l	1.0
14	Manganese	Mn	mg/l	0.5
15	Copper	Cu	mg/l	3.0
16	Zinc	Zn	mg/l	15.0
	Organic Pollution of Natural Origin			
17	BODs (5 days)	O_2	mg/l	6.0
18	PV (Oxygen abs. $KMnO_4$)	O_2	mg/l	20
19	Ammonium	NH_3	mg/l	2.0
20	Total Nitrogen Exclusive Nitrate		mg/l	1.0
	Organic Pollution Introduced Artificially			
21	Surfactants ABS (Alkyl Benxyl Sulphonates)		mg/l	2.0
22	Organic matter as carbon in chloroform extract)		mg/l	0.5
23	Phenolic substances as phenol		mg/l	0.002

Table 6.3: Radioactive Materials

Material	Limit
Gross alpha activity	0.1 Bq/l
Gross beta activity	0.1 Bq/l