Tables and figures

Table 1. Summary of evaluation options, with focus on main internal validity threats, relevance, and practical considerations that are of particular importance for network water supply and sanitation

@ >p(- 8) * >p(- 8) * >p(- 8) * >p(- 8) * >p(- 8) * @ **Method** & **Description and comments** & **Threats to validity of causal inference** & **Relevance of evaluation evidence** & **Practical / logistical considerations**

Experimental & & & &

Randomized Controlled Trial (RCT) [*Duflo et al.*, 2007] & RCTs are generally not feasible for network water infrastructure, as such interventions are clustered, directional, and designed to serve population at scale or to address known (selected) system deficiencies. Some complementary interventions (information campaigns) can be evaluated using this approach. Smaller-scale rural infrastructure (e.g., condominial sewerage, village-scale piped water) can be evaluated with cluster RCTs, or step-wedge RCTs. &

- *Confounding* due to unbalanced randomization
- *Spillovers* (violation of the stable unit treatment value assumption, or SUTVA), whereby some units benefit as a result of other units' uptake.
- Vulnerable to *selective attrition*

&

- Typically *artefactual*, w/ limited evaluation questions
- Treatment effect can be *representative*
- "Gold standard" for causal researchers
- Results are *not conditioned* by assumptions
- *Statistical power* is a design feature, but usually sufficient for a few pre-identified outcomes

&

- *Cost*: High, especially when powered for multiple outcomes or interventions
- *Contamination risk*: Moderate, as pressure to help "untreated" units increases over time
- *Coordination*: Mainly pertains to maintaining integrity of randomization
- *Interpretation*: Intuitive and highly transparent
- *Pre-intervention data needs*: Low to none

- *Flexibility to adapt*: Very low

Experimental encouragement design [*Katz et al.*, 2001] & Subsidies or other assistance to customers can generate exogenous variation in the take-up of infrastructure connections, for use as an instrumental variable for isolating impacts. The resulting local average treatment effect is specific to those who respond to the encouragement [*Heckman et al.*, 2006]. &

- Same as above

&

- Same as above, except that the treatment effect only applies to the population that responds to the encouragement

&
Quasi-experimental & & & &
Natural experiment [*J Angrist et al.*, 2002] & Some infrastructure placements are determined by geographic or other factors that are "as good as random" in determining exposure to improvements, such that they provide researchers with "natural experiments" [*Cerdá et al.*, 2012], that give rise to comparable treatment and control groups. Another version of this is an interrupted time series analysis where a time-dependent event (e.g., rehab of one part of a water network) gives rise to a sharp change that only affects some households or others. &

- *Confounding* by geographic / other factors determining exposure may also confound outcomes
- *Spillovers* (i.e., violation of SUTVA) outside of treatment area

&

- Evidence arises directly from the *real world*
- Treatment effect is *representative* but contingent on natural experiment conditions
- Generally accepted by researchers
- Results are *not conditioned* by assumptions
- *Statistical power*: Difficult to anticipate ex ante

&

- *Cost*: Low to moderate, depending on data collection needs
- *Contamination risk*: Low
- *Coordination*: Moderate; mainly in combining with other methods (DiD) to strengthen validity
- *Interpretation*: Intuitive but not always transparent

- *Pre-intervention data needs*: Low to none
- *Flexibility to adapt*: Impossible
- *Other*: Natural experiment can be hard to anticipate

Difference-in-differences (DiD) [*Card and Krueger*, 2000] & In this approach, impacts are estimated by subtracting out the trend in an unexposed sample, which represents the counterfactual, from that in an exposed sample. Such samples are created using variation in spatial targeting or other eligibility criteria, which are common for network water infrastructure extension or rehabilitation. The validity of the comparison relies on pre-treatment trends being similar in the groups, and can be enhanced using matching or econometric models that control for differences in baseline covariates. &

- *Confounding by* time-varying unobservables
- *Spillovers* (i.e., violation of SUTVA)
- Vulnerable to *selective attrition*

&

- Evidence arises directly from the *real world*
- Treatment effect is usually *representative* (unless combined w/other methods)
- Generally accepted by researchers, subject to showing parallel trends
- Results are *not conditioned* by assumptions
- *Statistical power* is a design feature

&

- *Cost*: Moderate to high, depending on data collection needs
- *Contamination risk*: Moderate to high
- *Coordination*: Moderate; mainly in combining with other methods (matching) to strengthen validity
- *Interpretation*: Intuitive and transparent
- *Pre-intervention data needs*: Moderate to high (parallel trends)
- *Flexibility to adapt*: Moderate

Matching or synthetic control [*Abadie and Gardeazabal*, 2003; *Rosenbaum and Rubin*, 1985] & These methods are best when combined with DiD analysis, but can be used to improve comparability when targeting is correlated with baseline characteristics. Various matching approaches enhance comparability

by sampling untreated observations that can approximate the treatment counterfactual. For example, propensity score matching (PSM) finds treated and untreated observations that have a similar probability of being treated, from a regression of participation on observables. Synthetic control uses a time series of pre-intervention observations to "train" an algorithm that identifies weights for a pool of observations with similar counterfactual trends as one or more treated units. &

- *Confounding* by unobservables (Conditional Independence Assumption), worse when match quality is low
- *Spillovers* (i.e., violation of SUTVA)

&

- Evidence arises directly from the *real world*
- Treatment effect only applies to units with suitable comparisons (common support region)
- Researchers are often skeptical that the CIA has been met
- Results are *conditioned* by assumptions of the matching algorithm
- *Statistical power* is a design feature

&

- *Cost*: Moderate to high, depending on data collection needs
- *Contamination risk*: High
- *Coordination*: Moderate; mainly in combining with other methods (DiD) to strengthen validity
- *Interpretation*: Intuitive, but matching may lack transparency
- *Pre-intervention data needs*: Moderate (matching)
- *Flexibility to adapt*: Moderate

Instrumental variables (IV) [*J D Angrist and Krueger*, 2001] & An instrumental variable is a factor that predicts exposure to or participation in an intervention, but that does not affect outcomes directly through channels other than that effect on participation. This creates exogenous variation in the intervention that can be leveraged to determine its impacts. The impact measure is a local average treatment effect that measures the effect of the intervention on those ("compliers") whose participation is affected by the instrument. Program placement rules or constraints may give rise to valid instruments. &

- *Confounding:* For many interventions and outcomes, there are few plausibly "exogenous" assignments of this type, at least in a statistical sense
- *Spillovers* (i.e., violation of SUTVA)

&

- Evidence arises directly from the *real world*
- Treatment effect (LATE) is not representative, and not always for the most relevant population
- Researchers are often skeptical about exclusion restriction
- Results are *conditioned* by exogeneity assumptions
- *Statistical power* is often reduced by 2-stage estimation

&

- *Cost*: Low to moderate, depending on data collection needs
- *Contamination risk*: Not applicable
- *Coordination*: Low
- *Interpretation*: Unintuitive, lacks transparency
- *Pre-intervention data needs*: Low
- *Flexibility to adapt*: High
- *Other*: Suitable IV may not exist

Regression discontinuity (RD) [*Imbens and Lemieux*, 2008; *Thistlethwaite and Campbell*, 1960] & RD exploits discontinuities in eligibility for an intervention with respect to an assignment variable. For example, population thresholds, or a poverty line threshold for subsidy eligibility. &

- *Confounding*: Eligibility rule violations or manipulation, or "fuzzy" discontinuities that are difficult to characterize well
- *Spillovers* (i.e., violation of SUTVA)
- Vulnerable to *selective attrition*

&

- Evidence arises directly from the *real world*
- Treatment effect is limited to units very near the discontinuity
- Generally accepted by researchers
- Results are *conditioned* on proximity to eligibility cutoff

- Statistical power may be limited

&

- *Cost*: Low to moderate, depending on data collection needs
- *Contamination risk*: Moderate, depending on rigor with which eligibility is assessed
- *Coordination*: Low
- *Interpretation*: Intuitive, but transparency may be lacking due to definition of the RD bandwidth
- *Pre-intervention data needs*: Low
- *Flexibility to adapt*: Low

Other & & & &
Ex post regression & Statistical comparison of treated and untreated units, with statistical control for observed differences between the groups. Also commonly called "observational" comparisons. &

- *Selection:* Units that participate are systematically different than those that do not
- *Confounding* by unobservables
- *Spillovers* (i.e., violation of SUTVA)

&

- Evidence arises directly from the *real world*
- Treatment effect is usually representative
- Causal researchers are typically highly skeptical of results
- Results are *conditioned* on controls
- *Statistical power*: Difficult to anticipate ex ante

&

- *Cost*: Low to moderate, depending on data collection needs
- *Contamination risk*: Not applicable
- *Coordination*: Low
- *Interpretation*: Intuitive, but transparency may be lacking (contingent on choice of controls)
- *Pre-intervention data needs*: None
- *Flexibility to adapt*: High

Counterfactual modeling [*Balke and Pearl*, 2013] & Complex water resources systems evolve stochastically according to both human and environmental influences. This approach leverages systems understanding from socio-hydrological or hydro-economic models to conduct "with" and "without" simulations of interventions, for construction of model-based comparisons [*Srinivasan*, 2015]. &

- *Confounding* by behavioral or other system-level factors not accounted for

&

- Evidence is *artefactual*; model may diverge from real world observations
- Treatment effect is usually representative, but may not align with policy-maker priorities and needs
- Not widely used by causal social science researchers, who are wary of over-calibration
- Results are *conditioned* on model assumptions
- *Statistical power*: Not applicable

&

- *Cost*: Low
- *Contamination risk*: Not applicable
- *Coordination*: Low
- *Interpretation*: Not intuitive and not always transparent (requires interdisciplinary expertise)
- *Pre-intervention data needs*: Moderate to high, depending on calibration needs
- *Flexibility to adapt*: High
- *Other*: Required model effort is substantial

Table 2. Summary of study populations and data collection methods deployed

@ >p(- 12) * >p(- 12) * >p(- 12) * >p(- 12) * >p(- 12) * >p(- 12) * >p(- 12) * @ **Survey element** & **Survey type** & **Sampling frame** & **Sample selection** & **Stratification / comparison group** & **Representation** & **Sample size**
Household & 4-wave Panel & Zarqa and Amman (from Jordan Dept. of Statistics (DoS)) &

- Survey geocodes selected based on *ex ante* matching of treated and control zones, using Census data
- Random sampling within sample geocodes
- Replacements selected from sample geocodes

&

WNP only – WNP only control

WWNP only – WWNP only control

WNP+WWNP – WNP+WWNP control

Distinct control groups from:

- Zarqa
- Amman

& Representative of sample geocodes at baseline, based on comparisons to Census and other sources &

1. 3359
2. 3416
3. 3596
4. 3662

Enterprise & 2-wave Panel & Zarqa and Amman (from DoS + household referrals) &

- Same geocodes as household sample
- Random selection within sample geocodes
- Referrals for informal enterprises
- Replacements selected from closely neighboring enterprises

& Same as household (though analysis uses all controls for each group to maximize statistical power) & Representative of sample geocodes at baseline

Informal enterprises likely under-represented (due to low referral rates) &

1. 345
2. 418

Farm & 3-wave Panel & Jordan Valley and highlands (from DoS) &

- Survey zones selected based on expected differences in exposure to treated wastewater

- Random selection in sample zones
- Replacements selected within zones

&

Five locations:

- Highlands u/s KTD (↑ river flow)
- JV1 North (↑ Non-Compact WW)
- JV2 Mid-North (↑ Compact WW)
- JV3 North-Central (↑ Compact WW)
- JV4 South-Central (little change in WW)

& Representative of sample zones &

1. 551
2. 539
3. 539

Refugee & Single cross-section & UNHCR registration list for Zarqa and Amman &

- Priority survey geocodes selected according to household sample, with augmenting based on treatment status outside hh geocodes
- Random sampling by treatment status
- Referrals for unregistered refugees

&

Treatment status:

- WNP only
- WWNP only
- Both WNP and WWNP
- Controls in Zarqa
- Control Amman

& Representative of registered population in sample areas

Unregistered population likely under-represented (due to low referral rates) & 1617
Water vendor & Single cross-section & Shops: Ministry of Health list + canvassing

Tankers: Canvassing &

- Full sampling from canvassed locations

& None & Representative of water vendors in Zarqa and East Amman in 2018 & 320
Meter testing & Repeat cross-section & Meter listing in selected zones &

- Zones selected for variation in JC status, elevation, pressure, throughput (for survey 1), and JC status and meter replacement (for survey 2)
- Random sample of meters within selected zones

& Compact and non-Compact zones & Not representative &

1. 37
2. 223

Water loss testing & Single cross-section & Canvassing of land plots in selected areas &

- "Well isolated" zones selected (as suggested by utility)
- Comparison of meter registered data to bulk meter inflow
- Random sub-sample of meters evaluated to adjust for meter error

& Meter error testing sub-sample stratified by meter replacement status & Not representative; only relevant to "well isolated" zones & 1797
Key informant interviews & Single cross-section & Listing of key JC stakeholders &

- Contact to all listed stakeholders
- Replacements included as suggested by stakeholders

& None & Representative of institutions, but likely not all perspectives & 22

Table 3. Summary of main impacts on household behaviors and outcomes

Outcome	**DiD impact of intervention – relative to non-intervention areas in Zarqa, by subsample**		**DiD impact of intervention – relative to non-intervention areas in Amman, by subsample**			
	(1) WNP	(2) WWNP	(3) Both	(4) WNP	(5) WWNP	(6) Both
Water supply						
Reported water pressure rating[1]	**-0.38*** (0.15)	n.a.	**-0.49*** (0.14)	**-0.63*** (0.12)	n.a.	**-0.84*** (0.14)
Reported perception of network water quality	n.a.			n.a.	(0.36)	
Assessed water quality (E. coli count)[2]	(0.053)	n.a.	(0)	n.a.	n.a.	n.a.
Hours piped water, for days w/water		n.a.	+0.37 (0.64)	**+1.15**** (0.46)	n.a.	**+1.83***** (0.47)
Reported water shortage, past month	**-0.10*** (0.05)	n.a.		**-0.12***** (0.04)	n.a.	

Outcome	DiD impact of intervention – relative to non-intervention areas in Zarqa, by subsample	DiD impact of intervention – relative to non-intervention areas in Amman, by subsample				
Network water use – Utility sample[3]	**+2.9*** (0.66)	n.a.	**+2.9*** (1.5)	+0.52 (0.59)	n.a.	**+1.9*** (1.1)
Network water use – Survey sample		n.a.	+6.1 (3.9)	+5.1 (3.5)	n.a.	+3.7 (4.3)
Expenditure on water from vendors (JD/month)	(5.0)	n.a.	(5.5)	(4.4)	n.a.	**-9.7*** (5.8)
Expenditure on water, all sources (JD/month)	(5.8)	n.a.	(5.7)	(5.1)	n.a.	(6.3)
Wastewater management						
Use of stand-alone cesspits	n.a.	**-0.13***** (0.04)	**-0.07*** (0.04)	n.a.	**-0.14***** (0.05)	**-0.11*** (0.04)
Sewer connection	n.a.	**+0.14***** (0.05)	**+0.17***** (0.05)	n.a.	**+0.09*** (0.05)	**+0.12**** (0.05)

Outcome	DiD impact of intervention – relative to non-intervention areas in Zarqa, by subsample			DiD impact of intervention – relative to non-intervention areas in Amman, by subsample		
Expense for septic tank evacuation	n.a.			n.a.		
Sewer backup prevalence	n.a.	(0.01)	(0.01)	n.a.	**-0.02** (0.01)**	(0.01)
Overall welfare						
Expenditure (JD/month)	(32.6)	+30.9 (39.5)	+15.6 (40.8)	+5.3 (32.8)	**+76.8*** (39.0)	(47.1)
Net income	(29.2)	(27.1)	(33.7)	**-66.6*** (38.7)	(35.8)	**-131***** (46.6)
Assets	+0.03 (0.03)	+0.04 (0.04)	+0.04 (0.04)	+0.04 (0.03)	+0.02 (0.04)	(0.04)
Sample size for comparison	,914	,443	,389	,359	,559	,418

Notes: All estimates are difference-in-differences estimates for coefficient $_t$ for period t=1 (after the intervention). Standard errors are shown in parentheses. Statistical significance is denoted as follows: *** p<0.01; ** p<0.05; * p<0.1. The specification controls for time and household fixed effects, and includes additional time-varying controls for the number of refugees arriving in the sample area (a demand shock) and a household-specific wealth index yield very similar

results (for alternative results omitting the controls and fixed effects, see Appendix 3). The subsample comparisons are as follows: WNP – Water Network Project treatment zones and matched control zones; WWNP – Wastewater Network Project treatment zones and matched control zones; Both – Water and Wastewater Network Project treatment zones and matched control zones.

[1] Measured on a four point scale (1 = excellent; 4 = poor)

[2] Water samples were only collected and analyzed in Zarqa

[3] The regressions for this outcome do not control for the time-varying factors because we use the full utility database, rather than restricting to the survey sample.

Table 4. Summary of main impacts on small enterprise behaviors and outcomes

Outcome	**DiD impact of intervention – relative to non-intervention areas in Zarqa, by subsample**		**DiD impact of intervention – relative to non-intervention areas in Amman, by subsample**			
	(1) WNP	(2) WWNP	(3) Both	(4) WNP	(5) WWNP	(6) Both
Water supply						
Piped water is primary source	(0.10)	n.a.	(0.10)	(0.11)	n.a.	+0.09 (0.12)
Hours piped water, for days w/water	(2.5)	n.a.	(2.5)	**-8.1*** ** (3.0)	n.a.	(3.0)
Water consumption (m^3/month)	+16.9 (16.8)	n.a.	+24.7 (19.2)	+12.5 (16.5)	n.a.	+25.9 (21.1)

Outcome	DiD impact of intervention – relative to non-intervention areas in Zarqa, by sub-sample	DiD impact of intervention – relative to non-intervention areas in Amman, by sub-sample				
Reported water interruption	+0.03 (0.11)	n.a.	(0.12)	+0.03 (0.14)	n.a.	+0.05 (0.15)
Expenditure on water from vendors (JD/month)	+0.21 (0.46)	n.a.	(0.49)	(0.54)	n.a.	**-1.50**** (0.61)
Expenditure on water, all sources (arcsin, JD/month)	+0.22 (0.28)	n.a.	(0.31)	**-0.57*** (0.33)	n.a.	(0.34)
Wastewater management						
Use of some wastewater system	n.a.	(0.08)	+0.02 (0.08)	n.a.	(0.10)	+0.05 (0.08)
Sewer connection	n.a.	**-0.18**** (0.09)	(0.09)	n.a.	(0.10)	+0.02 (0.09)

Outcome	DiD impact of intervention – relative to non-intervention areas in Zarqa, by subsample		DiD impact of intervention – relative to non-intervention areas in Amman, by subsample			
Cost of wastewater management (arcsin, JD/month)	n.a.	(0.32)	**-0.46*** (0.26)	n.a.	+0.05 (0.41)	(0.37)
Overall welfare						
Expenditure (arcsin, JD/month)	**0.46***** (0.15)	(0.15)	**+0.36*** (0.18)	**-0.34**** (0.17)		**-0.39*** (0.21)
Asset value (arcsin, JD)	+0.25 (0.33)	+0.09 (0.36)	**-0.77***** (0.28)	+0.24 (0.30)	(0.31)	(0.31)
Land value (arcsin, JD)	+0.57 (0.35)	+0.28 (0.45)	(0.37)	+0.60 (0.43)	+0.41 (0.53)	(0.34)
Sample size for comparison						

Notes: All estimates are difference-in-differences estimates for coefficient $_t$ for period t=1 (after the intervention). Standard errors are shown in parentheses. Statistical significance is denoted as follows: *** p<0.01; ** p<0.05; * p<0.1. The specification controls for time and enterprise fixed effects, as well as the following other time-varying factors: reported complaints about sewer overflows; respondent years with enterprise, number of total employees, reported obstacles

to growth, and frequency of water interruptions. Alternative specifications without controls yield very similar results (see Appendix C). The subsample comparisons are as follows: WNP – Water Network Project treatment zones and matched control zones; WWNP – Wastewater Network Project treatment zones and matched control zones; Both – Water and Wastewater Network Project treatment zones and matched control zones.

Table 5. Summary of main impacts on farm behaviors and outcomes

Outcome	**DiD impact of intervention – relative to non-intervention areas in Zarqa, by subsample**				
	(1) JV1	(2) JV2 (Treatment)	(3) JV3 (Treatment)	(4) JV4	(5) Highlands (Treatment)
Wastewater use in irrigation	**+0.10****	**+0.14*****	**+0.15*****	**-0.47****	**+0.08***
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Perceived water quality[1]	**-0.81*****	**-0.53***	(0.31)	**+1.40*****	(0.34)
	(0.3)	(0.31)		(0.31)	
Irrigated area (dunum)	**+6.0***	**+8.4****	**-9.0****	**-19.9*****	**+12.5*****
	(3.4)	(3.5)	(3.6)	(3.5)	(3.8)
Farm revenue (JD/yr)	**+91823***	(51173)	(51336)	**-91772***	**+186422*****
	(48757)			(51908)	(55180)
Farm profit (JD/yr)					
Farm land value (JD)					

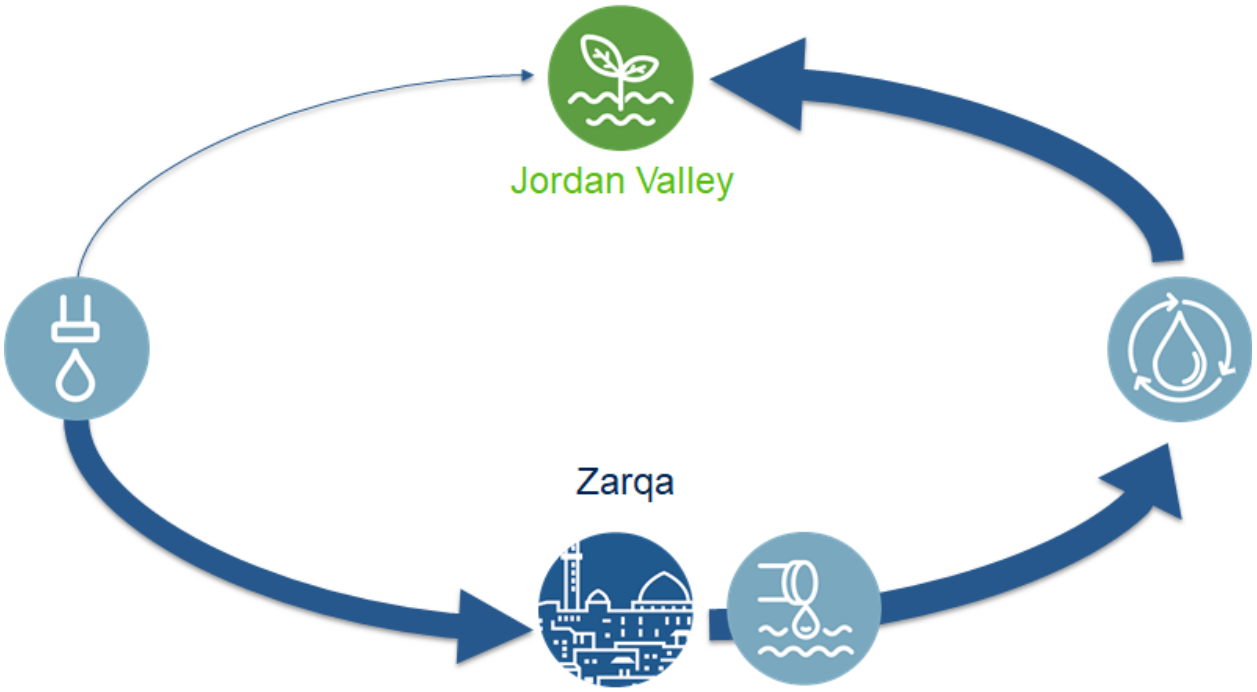
Notes: All estimates are difference-in-differences estimates for coefficient $_t$ for

period t=1 (after the intervention). Standard errors are shown in parentheses. Statistical significance is denoted as follows: *** p<0.01; ** p<0.05; * p<0.1. The specification controls for time and farm fixed effects. The subsamples are as follows: JV1 is furthest north in the Jordan Valley, and represents a set of farms that were mostly unaffected by the Compact since their water supply is independent of the Zarqa system; JV2 and JV3 represent areas where flows of recycled wastewater newly arrived (JV2) and increased substantially (JV3); JV4 represents an area that already had substantial flows of recycled water prior to the investment; Highlands farms, finally, are located along the Zarqa River and also received access to more steady water supply.

[1] Measured on a ten point scale (1 = poor; 10 = excellent)

Table 6. Summary of utility performance indicators, relative to other urban utilities in Jordan

Result	**Indicator**	**Utility**	**2008**	**20**
Reduced NRW	Indicator: Total losses			
	Total losses; network (%)	Aqaba	20.6	23
		Amman	40.1	38
		Zarqa	56.1	54
	Total losses; network (L / Subscriber / Day)	Aqaba	445	48
		Amman	340	30
		Zarqa	583	56
	Indicator: Pipe breaks/bursts per km of mainlines			
	Main bursts/ 100km	Aqaba	139	10
		Amman	109	89
		Zarqa	n.d.	n.
	Service leaks / 1000 connections	Aqaba	122	12
		Amman	221	20
		Zarqa	n.d.	n.
Increased revenue to utility	Utility revenue (2015 JD/m^3 sold)	Aqaba	1.08	0.
		Amman	1.40	1.
		Zarqa	0.90	0.
Increased cost recovery by utility	Billing efficiency (%)	Aqaba	91.5	91
		Amman	97.0	97
		Zarqa	n.d	n
	Collection efficiency (%)	Aqaba	99.4	10
		Amman	96.0	96
		Zarqa	n.d.	n.
	Operating Cost Recovery Ratio (OCRR)	Aqaba	1.43	1.
		Amman	1.07	1.
		Zarqa	0.87	0.



1. Qualitative depiction of (Top) pre- and (Bottom) ex
pected post- Jordan Compact situations

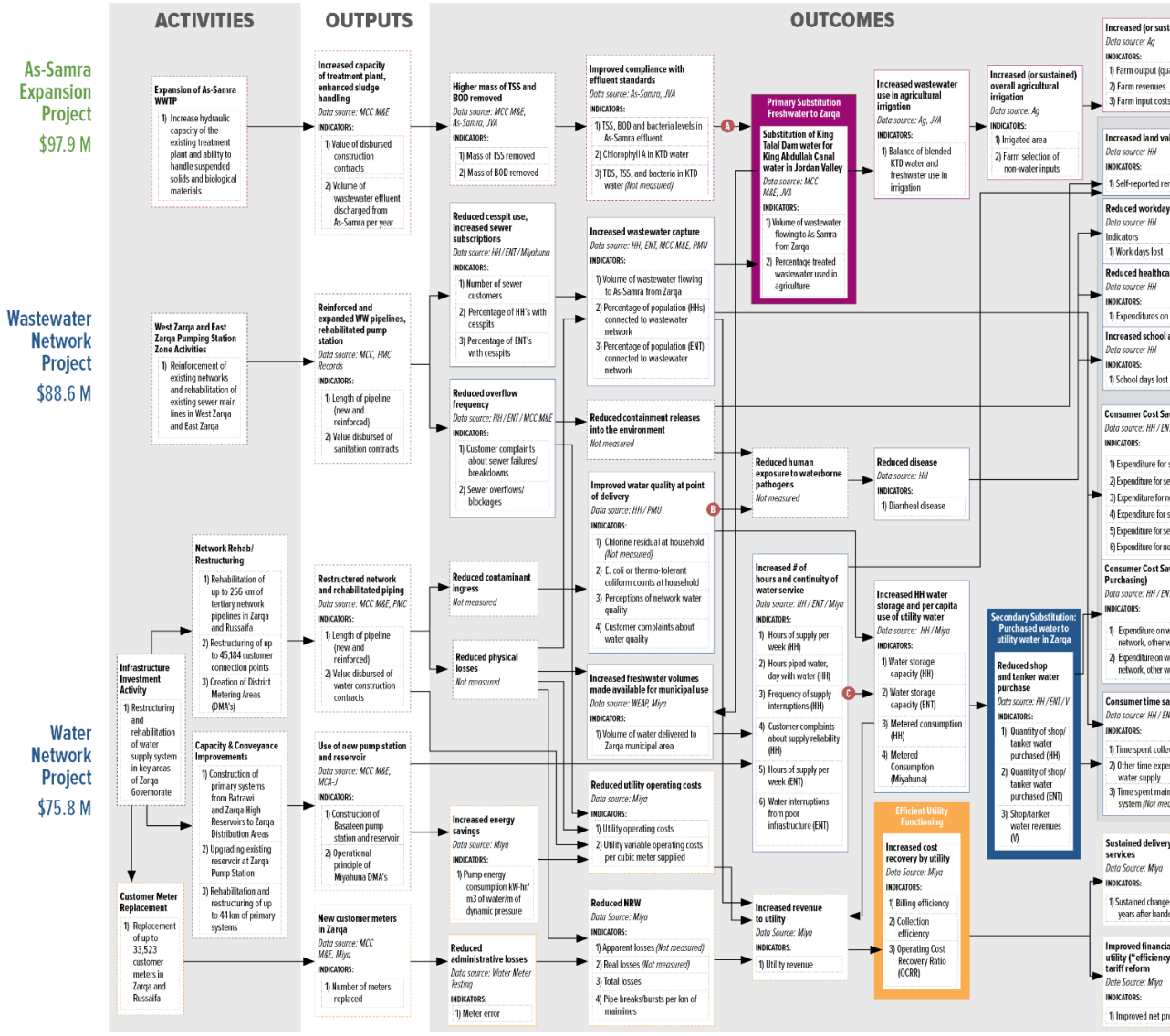


Figure 2. Full program theory of change, as elicited through participatory stakeholder consultations

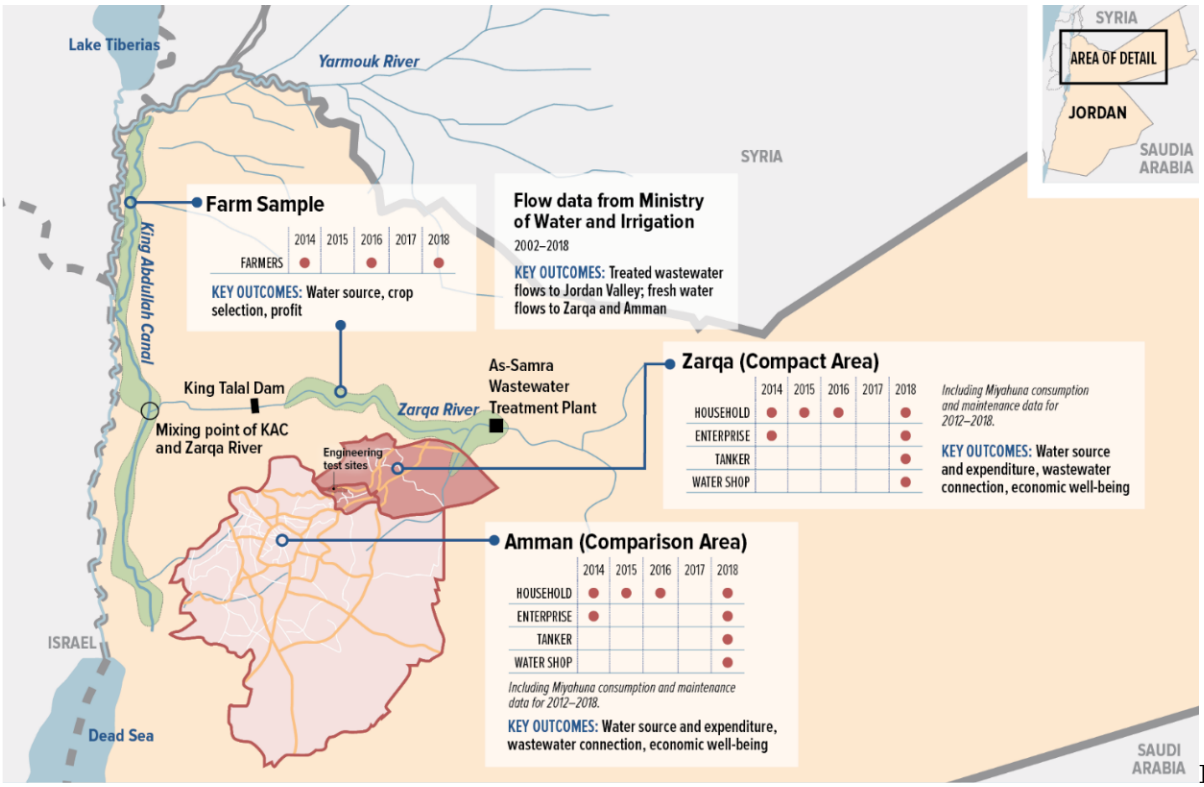


Figure 3. Locations and timing of data collection activities to support the evaluation

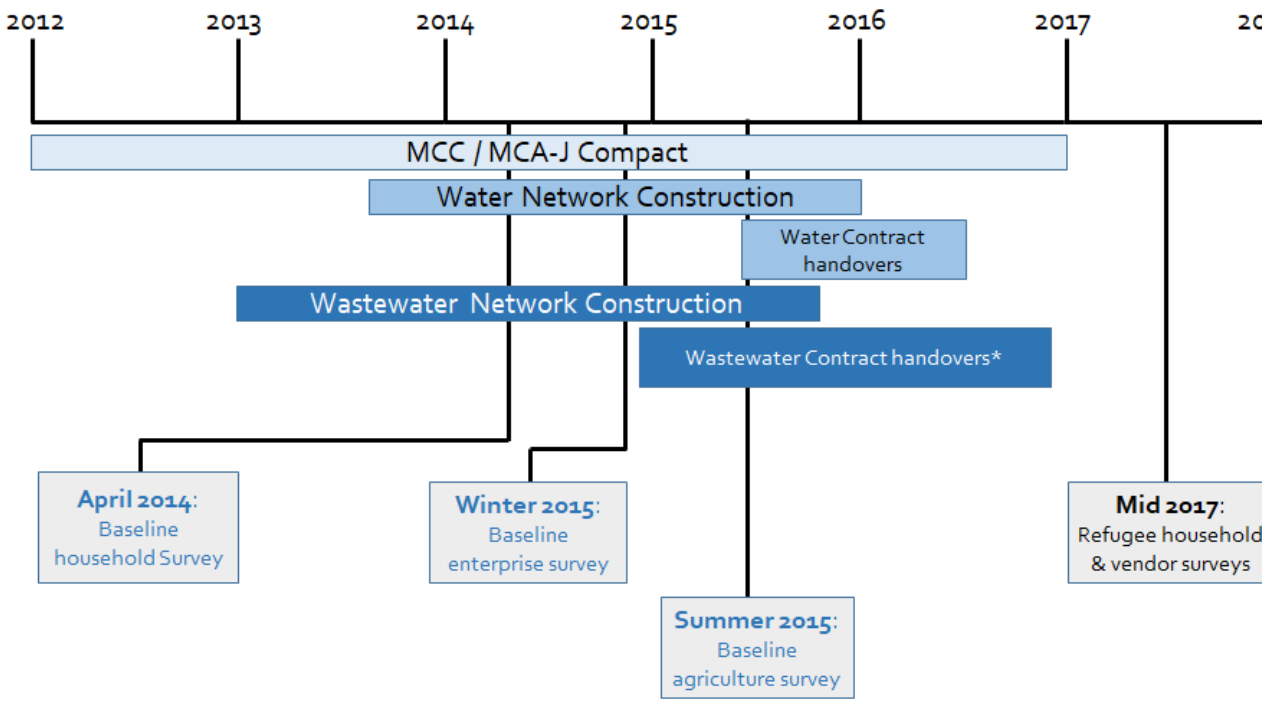


Figure 4. Timing of data collection relative to infrastructure intervention [Note that baseline surveys were conducted prior to any infrastructure operations, except for a few small wastewater Contract handovers preceding the baseline agriculture survey]