Lesotho – Evaluation of the Metolong Program and Urban and Peri-Urban Water Activity
Social Impact
Prepared on: April 6th, 2020 | *Revised May 29th, 2020*

MCC Evaluation Microdata
Data Package

Instructions

This template is informed by MCC's Evaluation Microdata Documentation and De-Identification Guidelines. In addition to reviewing these Guidelines, MCC contractors responsible for preparation and documentation of evaluation-related microdata for public and/or restricted-access use should be familiar with the following US government guidelines for data de-identification and re-identification:

- NIST 2015 - http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf
- NIST 2016 - http://csrc.nist.gov/publications/drafts/800-188/sp800_188_draft2.pdf

MCC, the evaluator, and stakeholders should consider the following multi-stage process for data review and release:

1. Evaluator and M&E PM should agree on expected DRB review date as early as possible to confirm. This should be scheduled at least one month before Evaluator's contract expires.
2. Evaluator should submit full package to M&E PM. The package includes:
    - One completed Section 1 of the DRB Data Package Worksheet for ALL data components (i.e. individual, household, and community data for one survey round are three data components with different risks)
    - One completed Section 2 & 3 for EACH data component
    - Datasets and code package(s)
    - Informed consent(s)
    - Questionnaire(s)
    - Most recent Metadata file (for Evaluation Catalog entry)
3. M&E PM should review Metadata and DRB Data Package Worksheet for clarity and completeness. This may require one round of revision based on the M&E PM requests for clarity and completeness.
4. Evaluator should submit full package to M&E PM. M&E PM and the M&E DRB members should establish a first-round review and feedback to the Evaluator on the proposed data de-identification process. This may require a second round of revision to the package.
5. Evaluator should submit full package to M&E PM for the confirmed MCC DRB review date at least 2 weeks prior to confirmed DRB review date.
6. If any feedback/revisions are required following MCC DRB review, Evaluator should revise and resubmit full package to M&E PM with documented responses to MCC DRB feedback to ensure timely virtual review and clearance of the full package. All final de-identification efforts and their impact on verification of analysis should be documented in the evaluator's Transparency Statement available on the Evaluation Catalog.

All red font text are instructions in the Worksheet and must be replaced with standard black font with the contractor's response.

Unless otherwise agreed with MCC, the final document will be made public to complement/underlie the contractor's Transparency Statement to document the data preparation and de-identification process required for the public and/or restricted-access microdata and any impact on the data for verifying evaluation analysis and broader data usability.

Section 1: Cover Sheet

Overview of Data Package

(*Instructions: Include a paragraph summarizing each data package component included in the package. For example, if the package includes household, individual, and community level data sets, please include a paragraph summarizing each of these three components, including information on the content and timing of the data collection*.)

Overview: All datasets submitted are products of a single quantitative household survey questionnaire. There was a single electronic questionnaire form for the survey for all households, with appropriate routing as relevant for households included in the impact evaluation versus customer survey, plus a supplementary form for entering *E. coli* results. Dataset 1 below contains the vast majority of the data produced by the questionnaire – while the *E. coli* results were entered into a separate form (because of the time needed for incubation before results are apparent), results have been fully merged into Dataset 1 and are submitted together. Datasets 2 through 5 below are the product of "repeat-group" features within the questionnaire, i.e. asking a set of questions for every sub-unit of a given type *within* a surveyed household. For example, the child illness module is repeated for every child under five within a household; the water collection module is repeated for every water source used by the household; etc. Dataset #1 was the main dataset used for analysis, and dataset #2 was used for the analysis of diarrheal illness. All others were used to construct variables. We submit them because there is richness in datasets 3 through 5 that might be of interest to others along replication of variable construction.

1. **Dataset #1 - Household Dataset** [Main analysis dataset]: The household survey dataset includes questions about household demographics, socioeconomic status, water source use, water collection, reliability, problems with water supply, coping behaviors such as water treatment and storage, water-related illnesses, and other variables. This dataset also includes the results of water quality testing for *E. Coli* and chlorine residual conducted at the household at the time of the survey.

2. **Dataset #2 - Child Illness** (Module I) [Analysis dataset for child diarrheal illness]: The child illness module includes child-level data regarding diarrheal illness in the last two weeks and associated care-seeking behaviors. Codes from the analysis do files can be used to merge relevant household-level information into this dataset in order to replicate the diarrheal illness analysis.

3. **Dataset #3 - Household Roster** (Module C) [Used for constructing variables]: The household roster module is used to construct various demographic covariates used in the evaluation's treatment and outcome models. The household survey was issued between April and July 2019, around four to five years after most of the MCC-funded infrastructure was commissioned.

4. **Datasets #4 and 5 - Water Collection** (Module E) [Used for constructing variables]: The water collection (#4) module contains information about each water source that the household reported using for any domestic purpose. The recall module contains the same questions with reference to pre-intervention time period, to reconstruct baseline for some key variables. The questionnaire looped through a set of questions for each water source within the household, asking about seasonality, frequency of collection, volume, round-trip time, and expenditures per unit. These modules are used to construct some of the main outcomes for analysis.

Complementary Data

(Instructions: Complementary data collection efforts are those efforts that complemented the data packages under review for de-identification, but do not necessarily require de-identification. The evaluator should list these data and provide a brief summary on how they connect to any data package components and affect the data package components' de-identification. For example, if the geospatial data for the project infrastructure is collected and will be publicly released, it should be listed in the complementary data collection efforts.)

We have not submitted any complementary data as part of this data package. Other complementary data used for our evaluation that are not submitted are listed below and in all cases the evaluation team does not have permission to share these datasets on their own accord publicly, so they are not submitted along with this package. Some variables extracted or constructed using the first two sources listed below are included in our household dataset, while any potential linkage variables have been removed or recoded (see Section 3). Any consequence for replicability of variable construction are further discussed in the Transparency Statement.

- Utility (WASCO) geospatial network data (sampling, analysis)
- Utility (WASCO) customer records (sampling, analysis)
- Lesotho National Development Corporation (for the industry PE analysis)
- National Bureau of Statistics, enumeration area shapefiles (sampling)

Qualitative data: Key informant interview and focus groups – as agreed with MCC during design stage, qualitative data will not be submitted to the IRB.

Data Package Folder Contents

(Instructions: Please list the Data Package Component File Name, and then include the File Names of each of the corresponding required documents [Metadata, Worksheet, Informed Consent, Questionnaire, Other docs]. Only one de-identification worksheet per survey is requested unless discussed.)

Data Package				
Component	**Worksheet**	**Informed Consent**	**Questionnaire(s)**	**Other Documents**
Dataset #1 hh_wq_combined_Deidentified.dta	Annex 05_Data Package Worksheet_LSO_MP_UPUW_Final.docx (This document)	Consent_Script.docx (Extracted from questionnaire form)	survey_final_printable_English.pdf survey_final_printable_Sesotho.pdf survey_final.xlsx wq_results_final.pdf wq_results_final.xlsx	hh_recode_results_log.txt 1_de_id_analysis_MASTER CONTROL.do cs_analysis.do household_vargen.do ie_a_analysis.do ie_b_analysis.do ie_matched_other_analysis.do ie_other_analysis.do post_merge_vargen.do
Dataset #2 illness_Deidentified.dta	As above	As above	survey_final_printable_English.pdf	illness_analysis.do illness_vargen.do
Dataset #3 hh_roster_Deidentified.dta	As above	As above	survey_final_printable_english.pdf	roster_recode_results_log.txt roster_vargen_and_collapse.do
Dataset #4 collection_Deidentified.dta	As above	As above	survey_final_printable_English.pdf	collection_vargen_and_collapse.do
Dataset #5 recall_collection_Deidentified.dta	As above	As above	survey_final_printable_English.pdf	recall_collection_vargen_and_collapse.do

Section 2: Dataset #1 (Household Data)

	Response		Discussion/Explanation
Data + Code Completeness	**Complete**	**COMPLETE** We are submitting the complete dataset including code that enables the generation of constructed variables.	*To be considered Complete: Dataset(s) must include ALL DATA COLLECTED in the survey, unless otherwise removed for de-identification purposes and described in this worksheet.* *The available data must allow new users to replicate evaluator analysis to the extent allowable by providing the full data set + analysis code. The constructed variables may also be included in a dataset, but if the dataset+code produces those variables, it is not necessary.* *To be considered Incomplete: The available data only provides a sub-section of data as produced by the survey and/or the constructed variables only. Incomplete data files are limited in terms of full verification of analysis and/or broad usability of data and must be justified.*
	Incomplete		
Data Round(s):	Baseline only	**ENDLINE ONLY** This is an *ex post* evaluation. The submission covers the only round of quantitative data collection completed for the evaluation. Some pre-intervention data is reconstructed through the questionnaire.	*MCC is willing to trade-off broad use of individual rounds for more consistent de-identification protocols across rounds of data. Therefore, unless there is specific demand for the baseline/interim only data, or contractual requirements, MCC prefers contractors to prepare all data rounds in one package.* *If one stage only – please (i) confirm demand and/or contractual justification and (ii) discuss how preparation and release of this data as presented to the DRB may affect future data round releases.* *If combination, please discuss if this file replaces any previously published datasets.*
	Interim only		
	Endline only		
	Combination of rounds		
Informed Consent and IRB	High restriction	**LOW RESTRICTION** Promises of confidentiality in the informed consent refer to identifiers, and language there allows for public posting of de-identified data. (See Consent_Script.docx).	*MCC assumes DIRECT identifiers are always removed from any public-use file. With this assumption: Please refer to the informed consent statement – does it require: High restriction: access to data that includes indirect identifiers is limited to the contractor only; Medium restriction: access to data that includes indirect identifiers is limited to the contractor and qualified researchers, including MCC; Low restriction: data with indirect identifiers may be made public.* *Please discuss how the promises of confidentiality in the informed consent informed de-identification*
	Medium restriction		
	Low restriction		

				efforts. Please include any additional guidance provided by the IRB as applicable.
Geographic Identifiers	**Town (Highest)**	See below.	**Identify**	*Please provide justification on the identification/de-identification/complete removal of specific geographic regions. De-identifying at a higher geographic level may support privacy protection, but it may also reduce data usability. Please provide justification for recommendation.*
	Township	See below.	**De-identify**	
	Enumeration Area (Lowest)	See below.	**Remove**	
Knowledge of Treatment	High risk	**LOW RISK** While it may be well known which sites were included in the Compact, a household's status as a WASCO customer in towns known to have been included in the study is not enough to enable re-identification after direct, linkage, and indirect identifiers have been removed from the data. The same is true of unconnected households from IE sites. In Maseru, it is not well known outside WASCO which areas are served by Metolong or not.		*In some cases, general knowledge of treatment areas and/or inclusion of a treatment variable can significantly increase re-identification risk depending on the population affected. Please provide assessment of this re-identification risk and recommendation if considered high/medium risk.*
	Medium risk			
	Low risk			
Publication Type	**Public-use only**	As described in Section 3, re-identification is low probability given measures taken and low risk. We recommend public use.		*Please state for this data package: will there be public-use data only, restricted-use data only, or both and provide justification as this relates to enabling verification of evaluation results and/or broad usability of the data.*
	Restricted-use only			
	Both			

Elaborated version of highlighted portion of table:

Geographic Identifiers	**Town (Highest)**	The population sizes of the towns is as follows, based on external sources: Butha Buthe 35108 Leribe 38558 Mafeteng 39754 Mapoteng 23926 Maseru 330760 Mazenod 19744 Mohales Hoek 40040 Mokhotlong 12940 Morija 7595 Qachas Nek 15917 Quthing 27314 Roma 13347 Semonkong 7856	**Identify** This variable is required in order to replicate parts of the analysis and substantively important disaggregations by town.

		TY 24257 However, we believe that the more relevant quantity for our study is the relatively smaller size of the *target* population (eligible types of households within each town). These include: Butha Buthe 2174 Leribe 2862 Mafeteng 5004 Mapoteng 1384 Maseru 41125 Mazenod 1939 Mohales Hoek 3185 Mokhotlong 1605 Morija 510 Qachas Nek 1396 Quthing 1351 Roma 1399 Semonkong 405 TY 4241	
	Township	Relevant in Maseru only, since Metolong-supplied (Design B T group) was defined (Y/N) by township. Avg. # eligible* customers** per township: 443 Avg. # eligible customers sampled per township: 14 Percent of eligible sampled from townships: Avg.: 6% Min: 1% Max: 100% *Connected before a certain date, which is stated in the public report. **WASCO customers sampled from WASCO database.	**De-identify (replace name with random number, as name is unnecessary for analysis)**

	Enumeration Area (Lowest)	Relevant in design A only, and only in eligible EAs (those that include households within 300m of network). The numbers below are results from SI's listing survey and subsequent sampling for the IE. T households (connected) are also theoretically included in the WASCO database, though this is not where they were sampled from for Design A. Avg. # HHs in Roma eligible EAs (total): 28 Avg. # potentially Eligible T households in those EAs from listing exercise: 12 Avg. # potentially Eligible C households in those EAs from listing exercise: 9 Avg. # HHs in Morija eligible EAs(total): 50 Avg. # potentially Eligible T households in those EAs from listing exercise: 17 Avg. # potentially Eligible C households in those EAs from listing exercise: 15 Avg. # HHs in Semonkong eligible EAs(total): 45 Avg. # potentially Eligible T households in those EAs from listing exercise: 18 Avg. # potentially Eligible C households in those EAs from listing exercise: 21 Survey was attempted with all potentially eligible households, 50-100% of which per EA were surveyed for the IE after confirming their eligibility.	**Remove**

Section 2: Datasets #2-5

Datasets 2 through 5 are from the same household survey. As described earlier they are simply the product of "repeat-group" survey questions that ask about sub-units within the household. Therefore, each of those datasets can be merged with the household data using the unique household key, and thus information for Section 2 is exactly the same across all submitted datasets. Note that while there are no geographic identifiers within datasets 2-5; as stated just above they can be merged and would then contain the same information. **Thus, the table below for datasets 2 through 5 is exactly the same as the table provided above for dataset #1.**

	Response		**Discussion/Explanation**
Data + Code Completeness	**Complete**	**COMPLETE** We are submitting the complete dataset including code that enables the generation of constructed variables.	*To be considered Complete: Dataset(s) must include ALL DATA COLLECTED in the survey, unless otherwise removed for de-identification purposes and described in this worksheet.* *The available data must allow new users to replicate evaluator analysis to the extent allowable by providing the full data set + analysis code. The constructed variables may also be included in a dataset, but if the dataset+code produces those variables, it is not necessary.* *To be considered Incomplete: The available data only provides a sub-section of data as produced by the survey and/or the constructed variables only. Incomplete data files are limited in terms of full verification of analysis and/or broad usability of data and must be justified.*
	Incomplete		
Data Round(s):	Baseline only	**ENDLINE ONLY** This is an *ex post* evaluation. The submission covers the only round of quantitative data collection completed for the evaluation. Some pre-intervention data is reconstructed through the questionnaire.	*MCC is willing to trade-off broad use of individual rounds for more consistent de-identification protocols across rounds of data. Therefore, unless there is specific demand for the baseline/interim only data, or contractual requirements, MCC prefers contractors to prepare all data rounds in one package.* *If one stage only – please (i) confirm demand and/or contractual justification and (ii) discuss how preparation and release of this data as presented to the DRB may affect future data round releases.* *If combination, please discuss if this file replaces any previously published datasets.*
	Interim only		
	Endline only		
	Combination of rounds		
Informed Consent and IRB	High restriction	**LOW RESTRICTION** Promises of confidentiality in the informed consent refer to identifiers, and language there	*MCC assumes DIRECT identifiers are always removed from any public-use file. With this assumption: Please refer to the informed consent statement – does it require: High restriction:*

	Medium restriction	allows for public posting of de-identified data. (See Consent_Script.docx).		*access to data that includes indirect identifiers is limited to the contractor only; Medium restriction: access to data that includes indirect identifiers is limited to the contractor and qualified researchers, including MCC; Low restriction: data with indirect identifiers may be made public.* *Please discuss how the promises of confidentiality in the informed consent informed de-identification efforts. Please include any additional guidance provided by the IRB as applicable.*
	Low restriction			
Geographic Identifiers	**Town (Highest)**	See below.	**Identify**	*Please provide justification on the identification/de-identification/complete removal of specific geographic regions. De-identifying at a higher geographic level may support privacy protection, but it may also reduce data usability. Please provide justification for recommendation.*
	Township	See below.	**De-identify**	
	Enumeration Area (Lowest)	See below.	**Remove**	
Knowledge of Treatment	High risk	**LOW RISK** While it may be well known which sites were included in the Compact, a household's status as a WASCO customer in towns known to have been included in the study is not enough to enable re-identification after direct, linkage, and indirect identifiers have been removed from the data. The same is true of unconnected households from IE sites. In Maseru, it is not well known outside WASCO which areas are served by Metolong or not.		*In some cases, general knowledge of treatment areas and/or inclusion of a treatment variable can significantly increase re-identification risk depending on the population affected. Please provide assessment of this re-identification risk and recommendation if considered high/medium risk.*
	Medium risk			
	Low risk			
Publication Type	**Public-use only**	As described in Section 3, re-identification is low probability given measures taken and low risk. We recommend public use.		*Please state for this data package: will there be public-use data only, restricted-use data only, or both and provide justification as this relates to enabling verification of evaluation results and/or broad usability of the data.*
	Restricted-use only			
	Both			

Elaborated version of highlighted portion of table:

Geographic Identifiers	**Town (Highest)**	The population sizes of the towns is as follows, based on external sources: Butha Buthe 35108 Leribe 38558 Mafeteng 39754 Mapoteng 23926 Maseru 330760 Mazenod 19744	**Identify** This variable is required in order to replicate parts of the analysis and substantively important disaggregations by town.

		Mohales Hoek 40040 Mokhotlong 12940 Morija 7595 Qachas Nek 15917 Quthing 27314 Roma 13347 Semonkong 7856 TY 24257 However, we believe that the more relevant quantity for our study is the relatively smaller size of the *target* population (eligible types of households within each town). These include: Butha Buthe 2174 Leribe 2862 Mafeteng 5004 Mapoteng 1384 Maseru 41125 Mazenod 1939 Mohales Hoek 3185 Mokhotlong 1605 Morija 510 Qachas Nek 1396 Quthing 1351 Roma 1399 Semonkong 405 TY 4241	
	Township	Relevant in Maseru only, since Metolong-supplied (Design B T group) was defined (Y/N) by township. Avg. # eligible* customers** per township: 443 Avg. # eligible customers sampled per township: 14 Percent of eligible sampled from townships: Avg.: 6% Min: 1% Max: 100% *Connected before a certain date, which is stated in the public report. **WASCO customers sampled from WASCO database.	**De-identify (replace name with random number, as name is unnecessary for analysis)**

	Enumeration Area (Lowest)	Relevant in design A only, and only in eligible EAs (those that include households within 300m of network). The numbers below are results from SI's listing survey and subsequent sampling for the IE. T households (connected) are also theoretically included in the WASCO database, though this is not where they were sampled from for Design A. Avg. # HHs in Roma eligible EAs (total): 28 Avg. # potentially Eligible T households in those EAs from listing exercise: 12 Avg. # potentially Eligible C households in those EAs from listing exercise: 9 Avg. # HHs in Morija eligible EAs(total): 50 Avg. # potentially Eligible T households in those EAs from listing exercise: 17 Avg. # potentially Eligible C households in those EAs from listing exercise: 15 Avg. # HHs in Semonkong eligible EAs(total): 45 Avg. # potentially Eligible T households in those EAs from listing exercise: 18 Avg. # potentially Eligible C households in those EAs from listing exercise: 21 Survey was attempted with all potentially eligible households, 50-100% of which per EA were surveyed for the IE after confirming their eligibility.	**Remove**

Section 3: Dataset #1 (Household Data)

Specific Issues		Risk Analysis		Risk Mitigation	
		Instructions	*Response*	*Instructions*	*Response*
1.	Who has significant financial, legal, cultural, or other incentives to re-identify survey respondents?	*List all potential threats*[1]	The only party with any potential, credible interest in re-identifying survey respondents could be the utility (WASCO).		
2.	What is the potential value to these intruders?	*List all uses (for example:* capture delinquent tax payments, or stigmatize the respondent)	However, the dataset does not have much value to the utility with regard to re-identifying individuals. The questionnaire asks about household characteristics, various aspects of water use in the household, diarrheal illness, and other topics pertinent to the research but otherwise unhelpful to the utility. If the utility were interested to re-identify those who re-sell water to neighbors, they might be able to do so if they had access to data with sufficient identifiers. However, the evaluation team understands that this is not an imminent risk (we understand existing enforcement is relatively lax), and with the measures taken below it would not be possible.		

[1] As stated in NIST 2016, de-identification practitioners should assume that de-identified US government datasets will be subjected to sustained, world-wide re-identification attempts, and they should gauge their de-identification requirements accordingly. Although a specific dataset may not be seen as sensitive, de-identifying that dataset may be an important step in de-identifying another dataset that is sensitive. Alternatively, the adversary may merely wish to embarrass the US government agency or its partners. Thus, adversaries may have a strong incentive to re-identify datasets that are seemingly innocuous.

Specific Issues		Risk Analysis		Risk Mitigation	
		Instructions	*Response*	*Instructions*	*Response*
3.	What is the expected cost to these intruders to re-identify the data?	*Describe degree of difficulty for re-identification*	The cost to re-identify is high. The party must decide that the time and effort required to potentially re-identify individual is less than alternative methods, such as going out into the community and investigating for themselves to identify any re-selling activity. The parties would have to have access to the appropriate software, or otherwise enlist separate individuals or institutions with the appropriate software, to access the data.		
4.	Assess availability of 'linkage' data that can be used to re-identify respondents. This includes other datasets or archives with information that can be used to re-identify individuals in the dataset.	*List all potential existing data*	<u>WASCO EDAMS database</u>: - **Account Number** - **Meter Number** - **Installation Date (mo./yr.)** <u>Household Survey</u> - **Bill information** was collected from some households, where a bill was available to observe (account, meter, relevant transaction amounts). Ultimately, few households had this information at the ready and the data was not useful for analysis. <u>Lesotho Bureau of Statistics</u>: - **Enumeration Area code:** Would connect a given household in Roma, Morija, or Semonkong, to its census enumeration area code	*Describe how to mitigate link to existing data that enables re-identification*	**All linkage variables are dropped from the dataset**

Specific Issues		Risk Analysis		Risk Mitigation	
		Instructions	*Response*	*Instructions*	*Response*
5.	**Identity Disclosures:** What are the DIRECT identifiers in the raw data?	*List the DIRECT identifiers* (names, addresses, geographic information, government-issued ID numbers, etc.)	Direct identifiers in the raw data include: - **Respondent Name** - **Head of household name** - **Respondent Phone number** - **Alternative phone number** - **Household GPS points** - **Geographic landmark** - **Household member names** Other identifiers: - Audio files and photos of doors were also collected from some households – the former randomly activated for quality control and the latter as part of listing to facilitate re-visits for the survey. **Neither of these is included in the data package. The dataset would have contained the file names associated with any households with such files.** The only places these file names would connect are with SI's internal records with those images and audio files. Tool • Note also that the name and phone number of the local data collection firm's field manager was written in the consent form. This has been redacted from the version of tools submitted.	*List all DIRECT identifiers removed from the dataset.*	**All direct identifiers have been dropped from the dataset.** **File names of audio files and door images have also been dropped.**
6.	**Attribute Disclosures:** For GIS/GPS data, this distance data can be a direct identifier that is VERY useful analytically. Therefore, please describe how GIS/GPS data VALUE/USABILITY can be retained.	*List all GPS and/or GIS data.*	The variable containing information about the **household's distance from the network** (meters) is required to replicate analysis. Household GPS points have been dropped as direct identifiers, while the distance variable has been retained. It is not sufficient alone or	*Describe process for de-identification. For example: introduce random errors into geographic data (GPS, GIS, etc.). Displace urban points 0-2 km, rural points 0-5 km, and*	**Distance to network (m) for each household has been retained**, while **GPS coordinates have been dropped** (see also item #5 above).

Specific Issues		Risk Analysis		Risk Mitigation	
		Instructions	*Response*	*Instructions*	*Response*
			with other variables to re-identify. Because eligibility was defined as within 300m, all distances are within this range.	*additional 1% of rural points 0-10 km^2.*	
7.	**Attribute Disclosures:** What variables have OUTLIERS that create INDIRECT identifiers are in the raw data?	*List the identifying items/variables*	With the removal of relevant direct identifiers and linkage variables, very few variables present any risk for re-identification. The evaluation team checked values for the following potentially sensitive or highly visible characteristics whose outliers might serve to identify: - Household (HH) ownership status - Head of HH age - Marital status - Highest level of education attained - Number of rooms used for sleeping - Number of HH members - Number of HH members under five - Observable household assets - Livestock - Number of tenants - Piped water consumption* - Water storage capacity - Household income With some exceptions, the outliers for most of these variables would not serve to indirectly identify a household. *Water consumption as collected in the household survey does not specifically link to	*Describe top/bottom coding: set upper & lower bounds to remove outliers for continuous. Specify: are values set to the median, or other? For large categories/datasets, the OMB suggests top coding at least the highest .5%; for smaller categories/datasets, top code the highest 3-5%. The same principles apply to bottom coding.*[3]	**Top-coded:** **Rooms used by household members for sleeping**: top coded 95th percentile: 236 of 4,668 observations top-coded as "4 or more" **Motor vehicles owned:** 109 of 4,667 observations top-coded as "3 or more" **Solar panels owned:** top-coded 95th percentile: 350 of 4,667 observations top-coded as "1 or more" **Motorcycles owned:** top-coded > 99th percentile: 33 of 4,668 observations top-coded as "1 or more". **Tractors owned:** top-coded 99th percentile: 85 of 4,668 observations top-coded as "1 or more."

[2] ICF International, Demographic & Health Surveys
[3] Office of Management and Budget, Checklist on Disclosure Potential of Proposed Data Releases (current link)

Specific Issues		Risk Analysis		Risk Mitigation	
		Instructions	*Response*	*Instructions*	*Response*
			outside datasets, since it includes water consumed from all sources (not just a tap if a household has one), and tap consumption does not come directly from a bill but rather from a question asking about average consumption. Nonetheless, we checked whether there might be any cases where households with taps reported such a large quantity of consumption that, within their town, using the EDAMS database, it could only be one of a small number of households with that level of consumption. Ultimately, we did not find any such cases.		**Refrigerators owned:** top-coded 95th percentile: 274 of 4,667 observations top-coded as "2 or more" **Hectares of land owned:** top-coded 99th percentile: 60 of 4,578 observations top-coded as "3 or more" **Cows owned:** top-coded 95th percentile: 29 of 480 observations top-coded as "7 or more." **Sheep owned:** top-coded 95th percentile: 24 of 475 observations top-coded as "36 or more." **Goats owned:** top-coded 95th percentile: 25 of 482 observations top-coded as "9 or more." **Horses owned:** top-coded 95th percentile: 39 of 487 observations top-coded as "1 or more." **Donkeys owned:** top-coded 95th percentile: 28 of 487 observations top-coded as "1 or more."

Specific Issues		Risk Analysis		Risk Mitigation	
		Instructions	*Response*	*Instructions*	*Response*
					Pigs owned: top-coded 95th percentile: 28 of 484 observations top-coded as "7 or more." **Chickens owned:** top-coded 95th percentile: 36 of 470 observations top-coded as "100 or more." **Head of household age:** top-coded 99th percentile: 49 of 4,353 observations top-coded as "86 or more." Note that a pre-constructed variable that is the square of this variable has also been recoded accordingly (i.e. all HoH who were 86 or older now have a squared HoH age of 7,396) **Tenants:** top-coded 95th percentile: 22 of 442 observations top-coded as 12 or more. **Number of storage containers of unusual size owned:** 47 of 4,104 observations top-coded as "5 or more" **Volume of storage containers of unusual size owned:** 9 of

Specific Issues		Risk Analysis		Risk Mitigation	
		Instructions	*Response*	*Instructions*	*Response*
					503 observations top-coded as "60 or more" liters The old and new variables are summarized in the recode log submitted with this worksheet.
				Describe any variables that require collapse and describe construction of new variable	**<u>Collapsed:</u>** New variables described below are more protective of identities and, moreover, more analytically useful (more meaningful and fewer categories). **HH ownership status:** Collapsed to own, rent, and other. **Head of HH marital status:** Collapsed granular categories of marital status with few observations into other relevant categories (e.g. living together or married, other). **Head of HH education:** Collapsed granular categories of education (e.g. by grade) into fewer and more meaningful categories (e.g. primary, secondary), etc.

Specific Issues		Risk Analysis		Risk Mitigation	
		Instructions	*Response*	*Instructions*	*Response*
					The old and new variables are cross-tabulated in the recode log submitted with this worksheet.
				Describe any global re-coding to group observations into categories (e.g., age 0-5, 5-10, 65+, etc.). Ensure that the categories are neither too broad nor too narrow.	N/A—where possible we have elected to preserve granular distribution of variables and top-code outlier values that might serve as indirect identifiers.
8.	**Attribute Disclosures:** What variable combinations produce UNIQUE observations that create INDIRECT IDENTIFIERS (for example: individuals with high incomes, ages, or unique combinations, such as 17-year old widowers or contextually unusual racial/ethnic backgrounds)	*List the identifying items/variables:*	After recoding the education and marital status variables and top-coding age variables, we are confident that other combinations of categorical or quantitative variables would not serve to produce unique and re-identifiable observations. Prior to recoding, there were such cases such as households of a certain size with a widow who is younger than a certain age. However, we have also redacted responses for some open text variables (including mainly text that follow selection of "other, specify" responses) that may serve to indirectly re-identify based on their mention of rare and conspicuous cases. For example, we redacted housing materials that were rare and highly visible, certain occupations	*For each identified rare data, describe the local suppression techniques employed to mitigate the identification risk of unique and rare observations. Specify: are values set to missing, the median, or other?*[4] *(See* **[Footnote]** *for MCC's general guidance; evaluators should either confirm that that this guidance is appropriate and was used, or explain the alternate method(s) used and why.)*	**<u>The following variables have had certain responses redacted</u>:** Other/specify text for: • fuel used for lighting • main material of floor • main material of roof • main material of walls • head of HH main activity • main motivation for connecting to WASCO • activities for which household does not have sufficient water • activities for which household did not have sufficient water previously

[4] To preserve the analytic value of rare data, MCC generally recommends replacing outlier values of continuous variables with the outlying group's median value – e.g., outliers in the 99th income percentile are replaced with the median of that quantile. And grouping rare categorical values with analytically similar categories (if meaningful similarities exist) or grouping them with other rare categories.

Specific Issues		Risk Analysis		Risk Mitigation	
		Instructions	*Response*	*Instructions*	*Response*
			that were rare and public-facing, or open text responses that revealed a household member with a sensitive status, like being disabled. In the data, those responses are now replaced with "REDACTED" In a small number of cases, we have dropped the entire variable when most or all responses could serve to re-identify households (such as **interviewer comments**, **other marital statuses**, **other household ownership statuses**, and **other income generating activities that use water**).		• Volume of unusually sized water storage containers • type of toilet household uses • reason household is not satisfied with toilet • method of disposing of child's stool • type of toilet household used before **The following variables were dropped entirely** to remove risk of re-identification: • Interviewer comments Other/Specify text for: • head of HH marital status • head of HH education • HH ownership status • income generating activities that use water

Section 3: Dataset #2 (Child Illness)

Specific Issues		Risk Analysis		Risk Mitigation	
		Instructions	*Response*	*Instructions*	*Response*
1.	Who has significant financial, legal, cultural, or other incentives to re-identify survey respondents?	*List all potential threats*[5]	The only party with any potential, credible interest in re-identifying survey respondents is the utility (WASCO). See Dataset #1. However, this does not pertain to information in the child illness dataset. We do not believe there are any other parties with interest to re-identify children in this dataset.		
2.	What is the potential value to these intruders?	*List all uses (for example:* capture delinquent tax payments, or stigmatize the respondent)	N/A We do not believe there are any other parties with interest to re-identify children in this dataset.		
3.	What is the expected cost to these intruders to re-identify the data?	*Describe degree of difficulty for re-identification*	The cost to re-identify is high. The parties would have to have access to the appropriate software, or otherwise enlist separate individuals or institutions with the appropriate software, to access the data. Further, the dataset asks about diarrheal illness in the two weeks prior to the survey and is therefore no longer actually observable to outsiders.		

[5] As stated in NIST 2016, de-identification practitioners should assume that de-identified US government datasets will be subjected to sustained, world-wide re-identification attempts, and they should gauge their de-identification requirements accordingly. Although a specific dataset may not be seen as sensitive, de-identifying that dataset may be an important step in de-identifying another dataset that is sensitive. Alternatively, the adversary may merely wish to embarrass the US government agency or its partners. Thus, adversaries may have a strong incentive to re-identify datasets that are seemingly innocuous.

Specific Issues		Risk Analysis		Risk Mitigation	
		Instructions	*Response*	*Instructions*	*Response*
4.	Assess availability of 'linkage' data that can be used to re-identify respondents. This includes other datasets or archives with information that can be used to re-identify individuals in the dataset.	*List all potential existing data*	There is no "linkage" data in the illness module.	*Describe how to mitigate link to existing data that enables re-identification*	**N/A**
5.	**Identity Disclosures:** What are the DIRECT identifiers in the raw data?	*List the DIRECT identifiers* (names, addresses, geographic information, government-issued ID numbers, etc.)	Direct identifiers in the raw data include: - **Household member name**	*List all DIRECT identifiers removed from the dataset.*	**Direct identifiers have been dropped.**
6.	**Attribute Disclosures:** For GIS/GPS data, this distance data can be a direct identifier that is VERY useful analytically. Therefore, please describe how GIS/GPS data VALUE/USABILITY can be retained.	*List all GPS and/or GIS data.*	N/A – no geographic variables in the illness module	*Describe process for de-identification. For example: introduce random errors into geographic data (GPS, GIS, etc.).* *Displace urban points 0-2 km, rural points 0-5 km, and additional 1% of rural points 0-10 km*[6]*.*	N/A – no geographic variables in the illness module
7.	**Attribute Disclosures:** What variables have OUTLIERS that create INDIRECT identifiers are in the raw data?	*List the identifying items/variables*	N/A – no such variables in the illness module.	*Describe top/bottom coding: set upper & lower bounds to remove outliers for continuous. Specify: are values set to the median, or other?* *For large categories/datasets, the*	N/A

[6] ICF International, Demographic & Health Surveys

Specific Issues		Risk Analysis		Risk Mitigation	
		Instructions	*Response*	*Instructions*	*Response*
				OMB suggests top coding at least the highest .5%; for smaller categories/datasets, top code the highest 3-5%. The same principles apply to bottom coding.[7]	
				Describe any variables that require collapse and describe construction of new variable	N/A
				Describe any global re-coding to group observations into categories (e.g., age 0-5, 5-10, 65+, etc.). Ensure that the categories are neither too broad nor too narrow.	N/A
8.	**Attribute Disclosures:** What variable combinations produce UNIQUE observations that create INDIRECT IDENTIFIERS (for example: individuals with high incomes, ages, or unique combinations, such as 17-year old	*List the identifying items/variables:*	N/A – no such variables in the illness module.	*For each identified rare data, describe the local suppression techniques employed to mitigate the identification risk of unique and rare observations. Specify: are values set to*	N/A

[7] Office of Management and Budget, Checklist on Disclosure Potential of Proposed Data Releases (current link)

Specific Issues		Risk Analysis		Risk Mitigation	
		Instructions	*Response*	*Instructions*	*Response*
	widowers or contextually unusual racial/ethnic backgrounds)			*missing, the median, or other?*[8] *(See* **[Footnote]** *for MCC's general guidance; evaluators should either confirm that that this guidance is appropriate and was used, or explain the alternate method(s) used and why.)*	

[8] To preserve the analytic value of rare data, MCC generally recommends replacing outlier values of continuous variables with the outlying group's median value – e.g., outliers in the 99th income percentile are replaced with the median of that quantile. And grouping rare categorical values with analytically similar categories (if meaningful similarities exist) or grouping them with other rare categories.

Section 3: Dataset #3 (Household Roster)

Specific Issues		Risk Analysis		Risk Mitigation	
		Instructions	*Response*	*Instructions*	*Response*
1.	Who has significant financial, legal, cultural, or other incentives to re-identify survey respondents?	*List all potential threats*[9]	The only party with any potential, credible interest in re-identifying survey respondents is the utility (WASCO). See Dataset #1. However, this does not pertain to information in the household member roster dataset. We do not believe there are any other parties with interest to re-identify individual household members in this dataset.		
2.	What is the potential value to these intruders?	*List all uses (for example:* capture delinquent tax payments, or stigmatize the respondent)	N/A We do not believe there are any other parties with interest to re-identify individual household members in this dataset.		
3.	What is the expected cost to these intruders to re-identify the data?	*Describe degree of difficulty for re-identification*	The cost to re-identify is high. The parties would have to have access to the appropriate software, or otherwise enlist separate individuals or institutions with the appropriate software, to access the data. The parties would have to spend a large amount of time trying to link common individual characteristics with those of people in corresponding town.		

[9] As stated in NIST 2016, de-identification practitioners should assume that de-identified US government datasets will be subjected to sustained, world-wide re-identification attempts, and they should gauge their de-identification requirements accordingly. Although a specific dataset may not be seen as sensitive, de-identifying that dataset may be an important step in de-identifying another dataset that is sensitive. Alternatively, the adversary may merely wish to embarrass the US government agency or its partners. Thus, adversaries may have a strong incentive to re-identify datasets that are seemingly innocuous.

Specific Issues		Risk Analysis		Risk Mitigation	
		Instructions	*Response*	*Instructions*	*Response*
4.	Assess availability of 'linkage' data that can be used to re-identify respondents. This includes other datasets or archives with information that can be used to re-identify individuals in the dataset.	*List all potential existing data*	There is no "linkage" data in the roster module.	*Describe how to mitigate link to existing data that enables re-identification*	**N/A**
5.	**Identity Disclosures:** What are the DIRECT identifiers in the raw data?	*List the DIRECT identifiers* (names, addresses, geographic information, government-issued ID numbers, etc.)	Direct identifiers in the raw data include: - Household member name	*List all DIRECT identifiers removed from the dataset.*	The following direct identifiers have been dropped from the de-identified dataset: - Household member name
6.	**Attribute Disclosures:** For GIS/GPS data, this distance data can be a direct identifier that is VERY useful analytically. Therefore, please describe how GIS/GPS data VALUE/USABILITY can be retained.	*List all GPS and/or GIS data.*	N/A – no geographic variables in the roster module	*Describe process for de-identification. For example: introduce random errors into geographic data (GPS, GIS, etc.).* *Displace urban points 0-2 km, rural points 0-5 km, and additional 1% of rural points 0-10 km[10].*	N/A – no geographic variables in the roster module
7.	**Attribute Disclosures:** What variables have OUTLIERS that create INDIRECT identifiers are in the raw data?	*List the identifying items/variables*	With the removal of relevant direct identifiers, very few variables present any risk for re-identification. The only exception is respondent age, which combined with other variables might serve to re-identify respondents. As was done for the head of household in component 1, we have top-coded age variables and collapsed/coarsened the	*Describe top/bottom coding: set upper & lower bounds to remove outliers for continuous. Specify: are values set to the median, or other?* *For large categories/datasets, the OMB suggests top coding at least the highest .5%; for*	**Household member age:** top-coded 99th percentile: 119 of 1,106 observations top-coded as "75 or more"

[10] ICF International, Demographic & Health Surveys

Specific Issues		Risk Analysis		Risk Mitigation	
		Instructions	*Response*	*Instructions*	*Response*
			education and marital status variables to mitigate the risk of re-identification.	*smaller categories/datasets, top code the highest 3-5%. The same principles apply to bottom coding.*[11]	
				Describe any variables that require collapse and describe construction of new variable	Household member marital status and education have been collapsed to new constructed variables that include all the granular categories collected in the survey, but in a variable more analytically useful (fewer categories) and more protective of identities. The construction of these variables is identical to what is done for the head of household variables in Component 1. Results of the change can be seen in the recode log.
				Describe any global re-coding to group observations into categories (e.g., age 0-5, 5-10, 65+, etc.). Ensure that the categories are neither too broad nor too narrow.	N/A
8.	**Attribute Disclosures:** What variable combinations produce	*List the identifying items/variables:*	After recoding the education and marital status variables and top-coding	*For each identified rare data, describe the local*	<u>The following variables have had certain responses redacted</u>:

[11] Office of Management and Budget, Checklist on Disclosure Potential of Proposed Data Releases (current link)

Specific Issues		Risk Analysis		Risk Mitigation	
		Instructions	*Response*	*Instructions*	*Response*
	UNIQUE observations that create INDIRECT IDENTIFIERS (for example: individuals with high incomes, ages, or unique combinations, such as 17-year old widowers or contextually unusual racial/ethnic backgrounds)		age variables, there are no more combinations of categorical or quantitative variables that would produce unique observations. Prior to recoding, there were such cases such as households of a certain size with a widow who is younger than a certain age. However, we have also redacted responses for some open text variables (including mainly text that follow selection of "other, specify" responses) that may serve to indirectly re-identify based on their mention of rare and conspicuous cases. For example, we redacted certain occupations that were rare and public-facing, or open text responses that revealed a household member with a sensitive status, like being disabled. In a small number of cases, we have dropped the entire variable when most or all responses could serve to re-identify households (such as other marital statuses and other education achieved).	*suppression techniques employed to mitigate the identification risk of unique and rare observations. Specify: are values set to missing, the median, or other?*[12] *(See* **[Footnote]** *for MCC's general guidance; evaluators should either confirm that that this guidance is appropriate and was used, or explain the alternate method(s) used and why.)*	Other/specify text for: • household member relationship to head of household • household member main activity **<u>The following variables were dropped entirely</u>** to remove risk of re-identification: Other/specify text for: 1.) household member marital status 2.) household member education

[12] To preserve the analytic value of rare data, MCC generally recommends replacing outlier values of continuous variables with the outlying group's median value – e.g., outliers in the 99th income percentile are replaced with the median of that quantile. And grouping rare categorical values with analytically similar categories (if meaningful similarities exist) or grouping them with other rare categories.

Section 3: Dataset #4 (Water Collection, current)

Specific Issues		Risk Analysis		Risk Mitigation	
		Instructions	*Response*	*Instructions*	*Response*
1.	Who has significant financial, legal, cultural, or other incentives to re-identify survey respondents?	*List all potential threats*[13]	The only party with any potential, credible interest in re-identifying survey respondents is the utility (WASCO). See Dataset #1. However, in large part we do not believe that this pertains to data within this particular dataset #4, because this dataset contains information about water collection from sources *other than one's own tap*. This contains information about those who collect from their neighbors' taps, but we believe WASCO would be far more interested in those who sell from their taps, rather than these households.		
2.	What is the potential value to these intruders?	*List all uses (for example:* capture delinquent tax payments, or stigmatize the respondent)	N/A We do not believe there are any other parties with interest to re-identify.		
3.	What is the expected cost to these intruders to re-identify the data?	*Describe degree of difficulty for re-identification*	The cost to re-identify is high. The parties would have to have access to the appropriate software, or otherwise enlist separate individuals or		

[13] As stated in NIST 2016, de-identification practitioners should assume that de-identified US government datasets will be subjected to sustained, world-wide re-identification attempts, and they should gauge their de-identification requirements accordingly. Although a specific dataset may not be seen as sensitive, de-identifying that dataset may be an important step in de-identifying another dataset that is sensitive. Alternatively, the adversary may merely wish to embarrass the US government agency or its partners. Thus, adversaries may have a strong incentive to re-identify datasets that are seemingly innocuous.

Specific Issues		Risk Analysis		Risk Mitigation	
		Instructions	*Response*	*Instructions*	*Response*
			institutions with the appropriate software, to access the data.		
4.	Assess availability of 'linkage' data that can be used to re-identify respondents. This includes other datasets or archives with information that can be used to re-identify individuals in the dataset.	*List all potential existing data*	There is no "linkage" data in the collection module.	*Describe how to mitigate link to existing data that enables re-identification*	**N/A**
5.	**Identity Disclosures:** What are the DIRECT identifiers in the raw data?	*List the DIRECT identifiers* (names, addresses, geographic information, government-issued ID numbers, etc.)	Direct identifiers in the raw data include: - Household member name (pulled from roster dataset dynamically by survey program)	*List all DIRECT identifiers removed from the dataset.*	The following direct identifiers have been dropped from the de-identified dataset: - Household member name (pulled from roster dataset dynamically by survey program)
6.	**Attribute Disclosures:** For GIS/GPS data, this distance data can be a direct identifier that is VERY useful analytically. Therefore, please describe how GIS/GPS data VALUE/USABILITY can be retained.	*List all GPS and/or GIS data.*	N/A – no geographic variables in the collection module	*Describe process for de-identification. For example: introduce random errors into geographic data (GPS, GIS, etc.).* *Displace urban points 0-2 km, rural points 0-5 km, and additional 1% of rural points 0-10 km[14].*	N/A – no geographic variables in the collection module

[14] ICF International, Demographic & Health Surveys

Specific Issues		Risk Analysis		Risk Mitigation	
		Instructions	*Response*	*Instructions*	*Response*
7.	**Attribute Disclosures:** What variables have OUTLIERS that create INDIRECT identifiers are in the raw data?	*List the identifying items/variables*	The value for quantitative variables in this dataset is critical for calculating evaluation outcome variables. They are also unlikely to be even indirectly identifying. Thus, we have left all as they are.	*Describe top/bottom coding: set upper & lower bounds to remove outliers for continuous. Specify: are values set to the median, or other? For large categories/datasets, the OMB suggests top coding at least the highest .5%; for smaller categories/datasets, top code the highest 3-5%. The same principles apply to bottom coding.*[15]	N/A
				Describe any variables that require collapse and describe construction of new variable	N/A
				Describe any global re-coding to group observations into categories (e.g., age 0-5, 5-10, 65+, etc.). Ensure that the categories are neither too broad nor too narrow.	N/A
8.	**Attribute Disclosures:** What variable combinations produce UNIQUE observations that create INDIRECT IDENTIFIERS (for	*List the identifying items/variables:*	Certain other/specify-sized collection containers might be conspicuously large when combined with other location and demographic variables.	*For each identified rare data, describe the local suppression techniques employed to mitigate the*	We have replaced any other/specify value over 500 liters with "Over 500 liters"

[15] Office of Management and Budget, Checklist on Disclosure Potential of Proposed Data Releases (current link)

Specific Issues		Risk Analysis		Risk Mitigation	
		Instructions	*Response*	*Instructions*	*Response*
	example: individuals with high incomes, ages, or unique combinations, such as 17-year old widowers or contextually unusual racial/ethnic backgrounds)			*identification risk of unique and rare observations. Specify: are values set to missing, the median, or other?*[16] *(See* **[Footnote]** *for MCC's general guidance; evaluators should either confirm that that this guidance is appropriate and was used, or explain the alternate method(s) used and why.)*	

[16] To preserve the analytic value of rare data, MCC generally recommends replacing outlier values of continuous variables with the outlying group's median value – e.g., outliers in the 99th income percentile are replaced with the median of that quantile. And grouping rare categorical values with analytically similar categories (if meaningful similarities exist) or grouping them with other rare categories.

Section 3: Dataset #5 (Water Collection, recall)

Specific Issues		Risk Analysis		Risk Mitigation	
		Instructions	*Response*	*Instructions*	*Response*
1.	Who has significant financial, legal, cultural, or other incentives to re-identify survey respondents?	*List all potential threats*[17]	The only party with any potential, credible interest in re-identifying survey respondents is the utility (WASCO). See Dataset #1. However, in large part we do not believe that this pertains to data within this particular dataset #4, because this dataset contains information about water collection from sources *other than one's own tap*. This contains information about those who collect from their neighbors' taps, but we believe WASCO would be far more interested in those who sell from their taps, rather than these households.		
2.	What is the potential value to these intruders?	*List all uses (for example:* capture delinquent tax payments, or stigmatize the respondent)	N/A We do not believe there are any other parties with interest to re-identify.		
3.	What is the expected cost to these intruders to re-identify the data?	*Describe degree of difficulty for re-identification*	The cost to re-identify is high. The parties would have to have access to the appropriate software, or otherwise enlist separate individuals or institutions with the appropriate		

[17] As stated in NIST 2016, de-identification practitioners should assume that de-identified US government datasets will be subjected to sustained, world-wide re-identification attempts, and they should gauge their de-identification requirements accordingly. Although a specific dataset may not be seen as sensitive, de-identifying that dataset may be an important step in de-identifying another dataset that is sensitive. Alternatively, the adversary may merely wish to embarrass the US government agency or its partners. Thus, adversaries may have a strong incentive to re-identify datasets that are seemingly innocuous.

Specific Issues		Risk Analysis		Risk Mitigation	
		Instructions	*Response*	*Instructions*	*Response*
			software, to access the data. In addition, this dataset pertains to collection behavior several years in the past.		
4.	Assess availability of 'linkage' data that can be used to re-identify respondents. This includes other datasets or archives with information that can be used to re-identify individuals in the dataset.	*List all potential existing data*	There is no "linkage" data in the recall collection module.	*Describe how to mitigate link to existing data that enables re-identification*	**N/A**
5.	**Identity Disclosures:** What are the DIRECT identifiers in the raw data?	*List the DIRECT identifiers* (names, addresses, geographic information, government-issued ID numbers, etc.)	Direct identifiers in the raw data include: - Household member name (pulled from roster dataset dynamically by survey program)	*List all DIRECT identifiers removed from the dataset.*	The following direct identifiers have been dropped from the de-identified dataset: - Household member name (pulled from roster dataset dynamically by survey program)
6.	**Attribute Disclosures:** For GIS/GPS data, this distance data can be a direct identifier that is VERY useful analytically. Therefore, please describe how GIS/GPS data VALUE/USABILITY can be retained.	*List all GPS and/or GIS data.*	N/A – no geographic variables in the recall collection module	*Describe process for de-identification. For example: introduce random errors into geographic data (GPS, GIS, etc.).* *Displace urban points 0-2 km, rural points 0-5 km, and additional 1% of rural points 0-10 km[18].*	N/A – no geographic variables in the recall collection module

[18] ICF International, Demographic & Health Surveys

Specific Issues		Risk Analysis		Risk Mitigation	
		Instructions	*Response*	*Instructions*	*Response*
7.	**Attribute Disclosures:** What variables have OUTLIERS that create INDIRECT identifiers are in the raw data?	*List the identifying items/variables*	The value for quantitative variables in this dataset is critical for calculating evaluation outcome variables. They are also unlikely to be even indirectly identifying. Thus, we have left all as they are.	*Describe top/bottom coding: set upper & lower bounds to remove outliers for continuous. Specify: are values set to the median, or other? For large categories/datasets, the OMB suggests top coding at least the highest .5%; for smaller categories/datasets, top code the highest 3-5%. The same principles apply to bottom coding.*[19]	N/A
				Describe any variables that require collapse and describe construction of new variable	N/A
				Describe any global re-coding to group observations into categories (e.g., age 0-5, 5-10, 65+, etc.). Ensure that the categories are neither too broad nor too narrow.	N/A
8.	**Attribute Disclosures:** What variable combinations produce UNIQUE observations that create INDIRECT IDENTIFIERS (for	*List the identifying items/variables:*	Certain other/specify-sized recall collection containers might be conspicuously large when combined	*For each identified rare data, describe the local suppression techniques employed to mitigate the*	We have replaced any other/specify value over 500 liters with "Over 500 liters"

[19] Office of Management and Budget, Checklist on Disclosure Potential of Proposed Data Releases (current link)

Specific Issues		Risk Analysis		Risk Mitigation	
		Instructions	*Response*	*Instructions*	*Response*
	example: individuals with high incomes, ages, or unique combinations, such as 17-year old widowers or contextually unusual racial/ethnic backgrounds)		with other location and demographic variables.	*identification risk of unique and rare observations. Specify: are values set to missing, the median, or other?*[20] *(See* **[Footnote]** *for MCC's general guidance; evaluators should either confirm that that this guidance is appropriate and was used, or explain the alternate method(s) used and why.)*	

[20] To preserve the analytic value of rare data, MCC generally recommends replacing outlier values of continuous variables with the outlying group's median value – e.g., outliers in the 99th income percentile are replaced with the median of that quantile. And grouping rare categorical values with analytically similar categories (if meaningful similarities exist) or grouping them with other rare categories.