**Consultants should complete the following *Anonymization Worksheet* and include it in the delivery of public use, anonymized data.**

The MCC Disclosure Review Board asks that the individuals responsible for anonymization prepare a response to the following questions in addition to this worksheet in preparation for the DRB review of the data:

1. **Understand Anonymization Process**: What was the general protocol/process followed for assessing risk and determining risk mitigation measures?

The Institutional Review Boards (IRBs) at both Stanford University and Virginia Tech reviewed and approved the research protocols. The development of the research protocols was an extensive process that required the Stanford-VT team to articulate the intent and scope of the research. For example, at Virginia Tech, the IRB protocol covered 15 different topics, including a justification of the research, how participants will be recruited into the study, the informed consent process, a description of the research procedures, an assessment of the risks and benefits associated with the research, the security of the data collected, and the confidentiality/anonymity of the final datasets. The process at Stanford University covered the same general subject areas. During the study, the IRB protocols were updated to take into account any changes that were made to the research design. For example, the Virginia Tech protocol was updated prior to the study of the small-scale solar systems that was added to the follow-up study.

1. **Understand Population Sizes**: What is the lowest geographic unit that can remain identified in the data based on risk assessment? What are the average population sizes by geographic unit?

The lowest identifiably unit in the dataset is an ‘anonymous’ household survey respondent. The table below documents the variables that have been removed from the datasets to ensure that survey respondents remain anonymous. Since the sampling of households was anchored around a handpump or the household of the community leader, there is no common geographic unit that can be used to calculate the average population sizes.

All community names and GPS coordinates have been removed from the datasets.

| **Country/Survey/Study Title:** *Impact Evaluation of the Mozambique Rural Water Supply Activity Under a Cooperative Agreement between MCC and Stanford University* | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Risk Level:** *Low* | | | | | |
| **Informed Consent Statement:** *The informed consent documentation for the household survey and water committee interviews has been submitted to the MCC. The institutional review boards at Stanford University and Virginia Tech have ruled that no identifying information collected during the 2011 baseline and 2013 follow-up studies can be submitted to the MCC. Thus, the final data sets submitted to the MCC are anonymized and can be released to the public.* | | | | | |
| **IRB Clearance:** *No* *personally identifiable information can be released to the public. Please see the above comment.* | | | | | |
|  | **Issues** | **Risk Analysis** | | **Risk Mitigation[[1]](#footnote-1)** | |
|  | | *Instructions* | *Response* | *Instructions* | *Response* |
| 1. | Individuals or organizations that might have significant financial, legal, cultural, or other incentives to de-anonymize survey respondents | *List all potential users* | Mozambique government;  Non-government organizations (e.g., World Vision); International development organizations (e.g., USAID, MCC, UNICEF, World Bank, etc.); Universities; National governments |  |  |
| 2. | Value or potential use of the data, if compromised. For example, some data might be used to capture delinquent tax payments, or could stigmatize the respondent | *List all uses* | The surveyor instruments were designed to minimize the potential risk associated with the questions being asked. If the data were compromised, individuals would be put at minimal risk since we did not collect information on sensitive issues such as political affiliations. |  |  |
| 3. | Cost of re-identification, both in terms of costs to potential intruders to de-anonymize data and the cost to potential victims of identification | *Describe degree of difficulty for re-identification* | Re-identifying individual respondents based on their responses will be very difficult and will require extensive background information on individuals in six districts within Nampula province. Households were select using a random walk, so there is no systematic way to identify the households participating in the study. Costs to potential victims of identification may include feeling that their privacy has been violated and that other people know more about their water collection, health, sanitation, feelings toward the water project, and income and expenditures than they have a right to know. |  |  |
| 4. | Direct & indirect identifiers (names, addresses, geographic information, government-issued ID numbers, etc.) | *List the identifying items/variables.* | Community and district names were recorded so that the same communities were included in the follow-up study.  The names of respondents and of children who were measured were recorded to ensure that households were re-interviewed during the follow-up study and children’s heights remeasured. | *List all identifying items/variables removed.* | The name of the **district** in which a household resides has been removed from the baseline (B003\_DISTRICT) and follow-up (F003\_DISTRICT) household survey codebooks.  The name of the **community** in which a household resides has been removed from the baseline (B007\_COMMUNITY) and follow-up (F004\_COMMUNITY) household survey codebooks.  The district and community names are indicated everywhere in the datasets by numbers rather than names.  All names have been removed from the following variables:  B018\_RESP\_NAME  B161\_CHILD1\_NAME  B166\_CHILD2\_NAME  F023\_RESP\_NAME  F037\_CHILD1\_REMEAS\_NAME  F043\_CHILD2\_REMEAS\_NAME  F051\_CHILD1\_NEWMEASURE\_NAME  F059\_CHILD2\_NEWMEASURE\_NAME  [Note: No identifying information is included in B181\_HOUSEOWN\_OTH. This question was not included in the follow-up study household survey.] |
| 5. | GIS/GPS data | *List all GPS and/or GIS data.* | GPS coordinates of the households surveyed and of community water points were recorded during the baseline and follow-up studies. | *Describe process for anonymization. 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[[2]](#footnote-2).* | The GPS coordinates of all the households surveyed have been removed from the following variables:  B564\_GPS\_LATITUDE  B564\_GPS\_LONGITUDE  F645\_GPS\_LATITUDE  F645\_GPS\_LONGITUDE  The GPS coordinates of the community water sources have been removed from the water testing data.  The GPS coordinates of all the households in which stored water was tested have been removed from the data sets. |
| 5. | Outliers in data that allow identification of respondent | *List the identifying items/variables:* | Remittances received may be potentially identifying information. | *Describe top/bottom coding: set upper & lower bounds to remove outliers for continuous.*  *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]](#footnote-3)* | All data on remittances have been removed from the following variables:  B506\_REMIT\_REL\_IND  B507\_REMIT\_REL\_2MONTH  B508\_REMIT\_NONREL\_IND  B509\_REMIT\_NONREL\_2MONTH  F581\_REMIT\_REL\_IND F582\_REMIT\_REL\_2MONTH  The top 5% of the following **income** variables in the baseline and follow-up datasets were top-coded:  B513\_AG\_CROP\_MONTH  B514\_AG\_WAGE\_MONTH  B519\_WAGE\_MONTH  B522\_LIVESTOCK\_YEAR  B527\_COMMERCE\_MONTH  B531\_SERVICE\_MONTH  B533\_FORESTPROD\_MONTH  B535\_HANDICRAFTS\_MONTH  B537\_DRINKS\_WEEK  B540\_FISHING\_WEEK  B543\_OTH\_MONTH  B545\_INCOME\_MONTH\_RESP  F585\_AG\_CROP\_MONTH  F586\_AG\_WAGE\_MONTH  F590\_WAGE\_MONTH  F592\_LIVESTOCK\_YEAR  F596\_COMMERCE\_MONTH  F597\_SERVICE\_MONTH  F598\_FORESTPROD\_MONTH  F599\_HANDICRAFTS\_MONTH  F600\_DRINKS\_WEEK  F602\_FISHING\_WEEK  F604\_OTH\_MONTH  F605\_INCOME\_MONTH\_RESP  F606\_INCOME\_YEAR\_RESP  The top 5% of the following **expenditure** variables in the baseline and follow-up datasets were top-coded:  B491\_EXP\_WEEK\_CURRY  B491\_EXP\_WEEK\_TOBACCO  B491\_EXP\_WEEK\_DRINKS  B491\_EXP\_WEEK\_SUGAR  B491\_EXP\_WEEK\_RICE  B491\_EXP\_WEEK\_LIGHTER  B491\_EXP\_WEEK\_BREAD  B491\_EXP\_WEEK\_SOAP  B491\_EXP\_WEEK\_CREDIT  B491\_EXP\_WEEK\_COOKOIL  B492\_EXP\_MONTH\_EMPLOYEE  B492\_EXP\_MONTH\_TRANSPORT  B492\_EXP\_MONTH\_ENERGY  B492\_EXP\_MONTH\_DOCTOR  B492\_EXP\_MONTH\_OINTMENT  B492\_EXP\_MONTH\_UTENSIL  B492\_EXP\_MONTH\_WATER  B493\_EXP\_YEAR\_CEREMONIES  B493\_EXP\_YEAR\_FEES  B493\_EXP\_YEAR\_CONSTRUCT  B493\_EXP\_YEAR\_CLOTHES  B493\_EXP\_YEAR\_ANIMALS  B496\_EXP\_OTH\_AMT  F564\_EXP\_WEEK\_CURRY  F564\_EXP\_WEEK\_TOBACCO  F564\_EXP\_WEEK\_DRINKS  F564\_EXP\_WEEK\_SUGAR  F564\_EXP\_WEEK\_LIGHTER  F564\_EXP\_WEEK\_BREAD  F564\_EXP\_WEEK\_SOAP  F564\_EXP\_WEEK\_CREDIT  F564\_EXP\_WEEK\_COOKOIL  F565\_EXP\_MONTH\_EMPLOYEE  F565\_EXP\_MONTH\_TRANSPORT  F565\_EXP\_MONTH\_ENERGY  F565\_EXP\_MONTH\_DOCTOR  F565\_EXP\_MONTH\_OINTMENT  F565\_EXP\_MONTH\_UTENSIL  F565\_EXP\_MONTH\_WATER  F565\_EXP\_MONTH\_RICE  F566\_EXP\_YEAR\_CEREMONIES  F566\_EXP\_YEAR\_FEES  F566\_EXP\_YEAR\_CONSTRUCT  F566\_EXP\_YEAR\_CLOTHES  F566\_EXP\_YEAR\_ANIMALS  F569\_EXP\_OTH\_AMT  The outliers for the following variables have been top-coded:  B524\_LIVESTOCK\_BULL   * 6 bulls or greater is top-coded to be 6   B524\_LIVESTOCK\_CHICKEN   * 20 chickens or greater is top-coded to be 20   B524\_LIVESTOCK\_PIGEON   * 9 pigeons or greater is top-coded to be 9   F593\_LIVESTOCK\_GOAT   * 10 goats or greater is top-coded to be 10   F593\_LIVESTOCK\_CHICKEN   * 20 chickens or greater is top-coded to be 20   F593\_LIVESTOCK\_DUCK   * 10 ducks or greater is top-coded to be 10   F593\_LIVESTOCK\_PIGEON   * 9 pigeons or greater is top-coded to be 9   B499\_RESP\_OWN\_HECT   * 4 hectares or greater is top-coded to be 4   F572\_RESP\_OWN\_HECT   * 4 hectares or greater is top-coded to be 4   F572\_RESP\_OWN\_LENGTH and F572\_RESP\_OWN\_WIDTH   * If the length times the width is greater than 4 hectares, both the length and width variables have been top-coded to 200   B036\_HHNUM\_ALL   * 9 people or greater is top-coded to be 9   F031\_HHNUM\_ALL   * 9 people or greater is top-coded to be 9   B448\_VISIT\_PROVHOSP\_NUM and  B451\_VISIT\_HEALTHPOST\_NUM   * 5 or greater is top-coded to be 5 * 99 = don’t know   B450\_VISIT\_PROVHOSP\_MONEY and  B453\_VISIT\_HEALTHPOST\_MONEY   * 300 or greater is top-coded to be 300   Since very few households engaged in agricultural employment the following variables have been set to NA in the baseline and follow-up datasets:  B501\_FARMEMPLOY\_IND  B502\_FARMEMPLOY\_NUM\_FULL  B502\_FARMEMPLOY\_NUM\_TEMP  B502\_FARMEMPLOY\_NUM\_TRADE  F575\_FARMEMPLOY\_IND  F576\_FARMEMPLOY\_NUM\_FULL  F576\_FARMEMPLOY\_NUM\_TEMP  F576\_FARMEMPLOY\_NUM\_TRADE  Since no data were collected in the follow-up study on the respondent’s participation in groups, the following variables have been recoded as NA in the baseline dataset:  B176\_PARTGROUPS\_a  B176\_PARTGROUPS\_b  B176\_PARTGROUPS\_c  B176\_PARTGROUPS\_d  B176\_PARTGROUPS\_e  B176\_PARTGROUPS\_f  B176\_PARTGROUPS\_g  B176\_PARTGROUPS\_h  B176\_PARTGROUPS\_i  B176\_PARTGROUPS\_j  B176\_PARTGROUPS\_k  B176\_PARTGROUPS\_l  B176\_PARTGROUPS\_m |
| *Describe any variables that require collapse and describe construction of new variable* | NA |
| *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.* | The religion variables in the baseline and follow-up datasets have been collapsed to the following three categories: 1 – Catholic; 2 – Muslim; and 3 – Other.  B178\_RELIGION  F069\_RELIGION  The asset variables in the baseline and follow-up datasets were recoded as follows: 1 – None; 2 – One or more of the following: Bike, Clock, Radio, Television, Car, Water tank larger than 200 liters, Barrow or animal tractor, Boat, Phone, Chairs, and/or Carpet.  B505\_ASSETS\_A  B505\_ASSETS\_B  B505\_ASSETS\_C  B505\_ASSETS\_D  B505\_ASSETS\_E  B505\_ASSETS\_F  B505\_ASSETS\_G  B505\_ASSETS\_H  B505\_ASSETS\_I  B505\_ASSETS\_J  B505\_ASSETS\_K  F579\_ASSETS\_A  F579\_ASSETS\_B  F579\_ASSETS\_C  F579\_ASSETS\_D  F579\_ASSETS\_E  F579\_ASSETS\_F  F579\_ASSETS\_G  F579\_ASSETS\_H  F579\_ASSETS\_I  F579\_ASSETS\_J  F579\_ASSETS\_K  The highest grade attained variables (listed below) in the baseline dataset have been collapsed as followed:  1 – None  2 – 1st to 6th grade  3 – 7th grade and above  99 – Don’t know  B020\_EDU1\_GRADE  B042\_EDU2\_GRADE  B057\_EDU3\_GRADE  B072\_EDU4\_GRADE  B087\_EDU5\_GRADE  B102\_EDU6\_GRADE  B117\_EDU7\_GRADE  B132\_EDU8\_GRADE  B147\_EDU9\_GRADE  The highest grade attained variables (listed below) in the follow-up dataset have been collapsed as followed:  1 – None  2 – Literate  2 – 1st to 6th grade  4 – 7th grade and above  99 – Don’t know  F025\_EDU1\_GRADE  F029\_SPOUSE\_EDU1 |
| 6. | Unique and rare observations, as well as highly visible variable combinations (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:* | Number of children under five years who died during the past two years could be a highly visible variable. | *For each identified rare data, describe the local suppression techniques employed to remove unique and rare data*  *Example: Replace observations within a unique/rare record with missing values (e.g., replacing ‘marital status = widow; age = 17’ with ‘marital status =. ; age = 17’)* | The data on child mortality have been removed from the following variables:  B546\_CHILD\_MORTAL  B547\_CHILD\_MORTAL\_CAUSE1A  B547\_CHILD\_MORTAL\_CAUSE1B  B547\_CHILD\_MORTAL\_CAUSE1C  B547\_CHILD\_MORTAL\_CAUSE1D  B548\_CHILD\_MORTAL\_CAUSE1OTH  B549\_CHILD\_MORTAL\_CAUSE2A  B549\_CHILD\_MORTAL\_CAUSE2B  B549\_CHILD\_MORTAL\_CAUSE2C  B549\_CHILD\_MORTAL\_CAUSE2D  B550\_CHILD\_MORTAL\_CAUSE2OTH  F608\_CHILD\_MORTAL  F609\_CHILD\_MORTAL\_CAUSE1A  F609\_CHILD\_MORTAL\_CAUSE1B  F609\_CHILD\_MORTAL\_CAUSE1C  F609\_CHILD\_MORTAL\_CAUSE1D  F610\_CHILD\_MORTAL\_CAUSE1OTH  F611\_CHILD\_MORTAL\_CAUSE2A  F611\_CHILD\_MORTAL\_CAUSE2B  F611\_CHILD\_MORTAL\_CAUSE2C  F611\_CHILD\_MORTAL\_CAUSE2D  F612\_CHILD\_MORTAL\_CAUSE2OTH |
| *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.* | NA |
| 7. | Availability of ‘key’ data that can be used to 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* | Household GPS coordinates and names of individuals will be retained by the Stanford/VT team. | *Describe how to mitigate link to existing data that enables re-identification* | These data are kept on password protected computers under the control of the Stanford/VT team. |
| 8. | Additional requirements by the Institutional Review Board(s) for data anonymization | *List all requirements* | No personally identifying information such as names and GPS coordinates may be released.  In addition, all community names and localities have been removed from the data sets. | *List responses to requirements* | No personally identifying information such as names and GPS coordinates have been released.  The names of all the communities surveyed in Nampula along with their associated localities have been removed from the datasets and from the final report. |

**Optional Section – Complete as necessary**

**Risk Mitigation – Data Perturbation**

1. For sensitive datasets, consider ‘micro-aggregation’ – average observations within *k* records to ensure that at least *k* units within a dataset are identical. For example, for the variable *age* and if *k*=4, take 4 observations, average their age, and replace each record’s *age* with the average.

Describe any micro-aggregation employed, and list the affected variables:

No micro-aggregation has been employed.

1. For sensitive data, consider ‘data-swapping’ – swapping values between records. This can be targeted (e.g., only unique records) or random; however, care must be taken not to alter the data too significantly.

Describe any data-swapping employed, and list the affected variables:

No data-swapping has been employed.

1. Add noise or resampling for selected variables. *Note that extensive data perturbation (i.e., perturbation of many variables) is not recommended as these processes can degrade data significantly*.

If noise was added or resampling occurred, please list the affected variables, why this procedure was deemed necessary and efforts taken to avoid degrading the data.

No noise was added to the data.

1. Any risk mitigation involving micro-aggregation, data swapping, adding noise or re-sampling should be documented as discussed in optional section below. [↑](#footnote-ref-1)
2. ICF International, Demographic & Health Surveys [↑](#footnote-ref-2)
3. Office of Management and Budget, Checklist on Disclosure Potential of Proposed Data Releases ([current link](https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&ved=0CCsQFjAA&url=http%3A%2F%2Fwww.fcsm.gov%2Fcommittees%2Fcdac%2Fcdac.html&ei=UN9vUpvxDZWt4APZyYD4Bg&usg=AFQjCNFwhGwSnNTfTDllVwmYgpJ2rdKEsg&bvm=bv.55123115,d.dmg)) [↑](#footnote-ref-3)