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
- NIST 2016

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*.)

This data package includes the following components:

The Millennium Challenge Corporation (MCC) invested in the Agriculture Development Project (ADP) as part of the Burkina Faso Compact. The project's objectives were to improve agricultural productivity, increase the incomes of farmers and livestock producers, and support economic development. The ADP was implemented from 2009 to 2014 and encompassed three activities: Water Management and Irrigation (WMI), Diversified Agriculture (DA), and Access to Rural Finance (ARF). Mathematica Policy Research was engaged by MCC as an independent evaluator to evaluate the WMI and DA activities.

This data package contains baseline data collected by the independent evaluators initially contracted to evaluate the ADP before Mathematica assumed its current role as evaluator. Mathematica did not use all baseline data delivered by MCC to Mathematica in the ADP baseline report submitted by Mathematica to MCC in Spring 2018. Mathematica determined that some baseline data were either unusable due to data quality concerns or unnecessary within Mathematica's approved evaluation design. As such, prior to the preparation of this data package, MCC and Mathematica discussed which baseline data should be delivered as part of the data package and in what form.[1] Below, we summarize the components of this data package by file type (restricted or public use):

Restricted-use data:

- Di PAP baseline survey data- The Di PAP baseline survey is a retrospective baseline survey. The survey was administered by BERD to a representative sample of 500 PAPs (out of roughly 1500) in October 2013 as some PAPs began receiving their new plots on the perimeter. A total of 388 PAPs out of the selected 500 PAPs completed the survey. The survey collected information on household demographics, production and land use outside the perimeter, perspectives on compensation, anticipated land use in the perimeter, household assets, access to credit, revenue, and training received. The survey data were received and are thus being delivered by Mathematica in 16 separate files.[2] The files map to specific sections of the baseline survey with each file at the observational unit level of the information collected. Mathematica used these data for the baseline report.
- Di Lottery baseline survey data- The Di Lottery baseline survey was administered in late 2013 by CERFODES to the 2,178 Di lottery applicants who met the lottery eligibility criteria (though not necessarily admitted to the lottery). A total of 2,128 applicants completed

[1] Mathematica's data delivery proposal was submitted to MCC in <<Burkina public use memo to MCC_revised_final_updated June 27 2019>>, which is included as an attachment to the data delivery package. Since the approval of the data delivery proposal, Mathematica and MCC determined that the interim crop yield data would not be submitted in a restricted- or public-use file because no GPS data were provided in the raw data submitted to Mathematica despite GPS coordinate fields appearing on the interim crop yield survey instrument.

[2] The codebooks for each separate data file have been collated into a single PDF file.

the survey. The survey collected data on demographic characteristics, socioeconomic status, agricultural experience, and other background characteristics relevant to the criteria for admission to the Di lottery. These survey data are delivered in a single data file in which the unit of observation is the lottery applicant. Mathematica used these data for the baseline report.

- Farmer training baseline household survey data- NORC and CERFODES designed and administered the farmer training baseline household survey which was to provide a baseline for the evaluation of a range of ADP activities. The baseline data were collected in two rounds in parallel with the two agricultural seasons in Burkina Faso. The first round of the baseline data was collected immediately after the 2011 dry season, and the second round was collected immediately after the 2011 rainy season. The survey's targeted sample comprised 1,082 matched pairs of farming households with each pair containing one household from the project's treatment area and the other from the comparison area. The lengthy baseline survey is comprised of seven modules focusing on the following content areas: household, agriculture, animal husbandry, forestry, consumption and credit, food security, and health. The data collected via each module was delivered to Mathematica in multiple data files given that the data collected under each module could have been collected at multiple levels (e.g. household, individual, plot, and crop levels). As such, we are delivering each data file as a unique restricted-use file (roughly 50 files per survey round/season) with each file at the observational unit level of the information collected.[3] Mathematica used a subsample of these data for the baseline report.[4]
- Barymetric survey data- The barymetric survey was administered by CERFODES to a subsample of farmer training households in two rounds: baseline in mid-2012 and one-year follow-up in mid-2013. In total, 153 households completed the baseline survey of which 146 completed the follow-up survey. In each round, the survey obtained data on cattle herd size, health, weight, milk production, and other related bovine information from each sampled household. The data for each round were received, and are thus delivered, in three separate files mapping to the three distinct sections of the survey: i) cattle herd characteristics at the household level; ii) cattle weight and other bovine characteristics at the cattle level; and iii) milk production at the cattle level.[5] Mathematica did not use these data in the baseline report.
- Farmer training supplemental household survey data- Implemented by CERFODES in late-2013, the farmer training supplemental household survey primarily collected information on the training and support farmers in the project's treatment area received from AD10. A short additional section of the survey also collected estimates of 2013 household agricultural production. Of the 1,082 farmer training households in the treatment area, 949 completed the survey. The data were received, and are thus delivered, in two separate files which map to the two sections of the survey: the first section covering training and support at the household level, and the second section covering household production at the crop level.[6] Mathematica did not use these data for the baseline report.

[3] The codebooks for each separate data file have been collated into a single PDF file for each agricultural season.
[4] Mathematica's approved evaluation design for the farmer training household data is a pre-post evaluation that only includes the 624 treatment households that received farmer training according to AD10. The farmer training baseline household restricted-use files contain all surveyed households (N=2,164). We include in this data package an Excel list of the 624 trained households' randomized IDs so that the user may anonymously identify them in the restricted-use data.
[5] The codebooks for each separate data file have been collated into a single PDF file for each data collection round.
[6] The codebooks for each separate data file have been collated into a single PDF file.

*Note: Mathematica prepared the restricted-use data files using cleaned baseline survey data. Mathematica applies the missing value codes listed in Table 1 during its data cleaning processes:

Table 1- Classifications of missing value codes used in Mathematica's data cleaning processes

.m	Missing or not ascertained (item nonresponse), referring to items that were skipped but should have been answered.
.e	Illogically complete, items should have been skipped but have been answered.
.s	Logical Skip (item nonresponse), referring to an item that was legitimately skipped based on prior (screener or filter) responses or on conditions of who is and who is not to answer a question or question set.
.n / .x	Not Applicable (item or unit nonresponse), including other reasons why a data item is not applicable to the case. Value code .n is used for individual survey questions that do not apply to an observation within a set of survey questions that generally do apply to the observation. Value code .x is used for an entire set of survey questions not applicable to an observation (e.g. an entire survey module not applicable).
.p	Processing Error. For some reason, there is no answer to the question (although the subject may have provided one). This can result from interviewer error, incorrect coding, machine failure, or other problems at the time of data entry.
.d	For responses marked as 'don't know' or 'don't remember'.
.r	Refused to answer the question.

Public-use data:

- Farmer training baseline household survey key indicator public-use file- As discussed with and approved by MCC, the public-use file we are submitting for the farmer training household survey contains key farmer training indicators created from the cleaned survey data. The key indicators contained in the public use file include type of irrigation used; use of improved seeds, organic and inorganic fertilizer, and pesticide; ADP training and adaptation of focus practices; yields (per hectare) by focus crop; and total agricultural profits (cfa). The file includes these indicators from both agricultural seasons, dry and rainy seasons 2011, and observations are unique to each household. Unlike the farmer training household restricted-use files which contain all surveyed households, the farmer training household key indicator public-use file only contains the 624 households from the project's treatment area that received farmer training according to AD10.

- Barymetric surveys key indicator public-use files- Similar to the above, the public-use files we are submitting for the two barymetric surveys (baseline in mid-2012 and one-year follow-up in mid-2013) contain key barymetric indicators created from the cleaned survey data. These indicators include cattle ID, breed, age, and gender; barymetric measurements (e.g. cattle height and weight); milk production measurements (e.g. liters of milk per day); and ADP training receipt (follow-up only). Because some cattle IDs are missing at follow-up or are inconsistent across rounds making merging data across rounds imperfect, we are submitting two barymetric key indicator public-use files, one for the baseline survey and one for the follow-up survey. The observations of each file are unique to each cow observed of each household sampled (153 households at baseline and 146 households at follow-up).

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*.)

This data package considers the following complementary data efforts:

Mathematica did not carry out any complementary data collection efforts related to the ADP baseline data submitted in this data package.

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.)*

Table 2: Data Package Components

Data Package				
Component	**Worksheet**	**Informed Consent**	**Questionnaire[7]**	**Other Documents**
Di PAP baseline survey RUFs[8]	BurkinaFaso ADP BL_DRB Data Package - Di PAP Worksheet.docx	Mathematica did not receive a consent statement for this survey. The survey questionnaire itself states that information will be "strictly confidential and only used for research purposes"	Di PAP Final Questionnaire- French.pdf	None
Di Lottery baseline survey RUF	BurkinaFaso ADP BL_DRB Data Package - Di Lottery Worksheet.docx	Page 1 of Questionnaire. Di Lottery baseline informed consent.pdf	Final Di Non-PAP Baseline Questionnaire- French.docx	None
Farmer training baseline household survey RUFs	BurkinaFaso ADP BL_DRB Data Package - FT HH Worksheet.docx	Farmer training household survey consent rainy season.pdf No consent renewal for dry season.	Dry season: • Agriculture Questionnaire- French.doc • Animal Husbandry Questionnaire- French.doc • Expense and Credit Questionnaire- French.doc • Food Security Questionnaire- French.doc	None

[7] For some data components, Mathematica did not receive complete survey instruments from MCC (we received what MCA had available). For example, we did not receive the Forestry and Consumption and Credit modules for the farmer training baseline household rainy season survey. We used the dry season version of these modules when cleaning the farmer training baseline household rainy season data. We also received only the milk production section of the barymetric follow-up survey. We used the barymetric baseline survey to facilitate the cleaning of the follow-up barymetric data.

[8] RUF= restricted-use file

			• Forestry Questionnaire- French.doc • Health Questionnaire- French.doc • Household Questionnaire- French.doc • Price Questionnaire- French.docx Rainy season: • Agricultural Questionnaire- French.doc • Animal Husbandry Questionnaire- French.doc • Food Security Questionnaire- French.doc • Health Questionnaire- French.doc • Household Questionnaire- French.doc	
Barymetric survey data RUFs	BurkinaFaso ADP BL_DRB Data Package - Barymetric Worksheet.docx	Mathematica was not provided a consent form.	Baseline: • questionnaire-barymetric baseline-bfa-agdev-dec14.pdf Follow-up: • questionnaire-barymetric milk followup-bfa-agdev-dec14.pdf	None
Farmer training supplemental household survey RUFs	BurkinaFaso ADP BL_DRB Data Package - FT Supp HH Worksheet.docx	Mathematica was not provided a consent form.	questionnaire-cropyields supplemental-bfa-agdev-nov13.doc	None
Farmer training baseline household survey key indicator PUF[9]	*see row "Farmer training baseline household survey RUFs"	*see row "Farmer training baseline household survey RUFs"	*see row "Farmer training baseline household survey RUFs"	None
Barymetric surveys key indicator PUFs	*see row "Barymetric survey data RUFs"	*see row "Barymetric survey data RUFs"	*see row "Barymetric survey data RUFs"	None

[9] PUF= public-use file

Section 2: Data Component Preparation Overview- Barymetric survey data

	Response		Discussion/Explanation
Data + Code Completeness	Complete	Complete	*To be considered Complete: 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	Baseline and one-year follow-up	*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.*
	Interim only		
	Endline only		
	Combination of rounds		

				If combination, please discuss if this file replaces any previously published datasets.
Informed Consent and IRB	High restriction	Medium restriction: We received IRB guidance that data can be submitted to ICPSR, and any institution that meets strict ICPSR requirements can access restricted use data. IRB recommends restricted use access with at least a secure download restriction.		*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 efforts. Please include any additional guidance provided by the IRB as applicable.*
	Medium restriction			
	Low restriction			
Geographic Identifiers	Highest (~~i.e. Province~~) Region	Population size (household level): 153	Identify: There are six regions represented in the data. De-identifying region would reduce the usability of the data given that region is the highest geographic level represented in the data and thus the most useful geographic identifier for analyses by specific areas of the country for which knowing specific region names would be important. However, we decided to combine the regions of the comparison areas since they contain few households: Centre (3 households) combined with Centre Nord (19), Plateau Central (11), and Hauts-Bassins (3). Boucle du Mouhoun (93) and Cascades (24) remain their own regions. As such, we 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.*

			region in the data but combine the comparison regions to reduce the likelihood of re-identification.	
	~~- (i.e. District)~~ Province	Avg. pop size (household level): 15	<u>De-identify</u>: There are 10 unique provinces represented in the data of which five have fewer than 10 households which could be identifying. Instead of combining the provinces with fewer than 10 households, we decided to randomize all province IDs because knowing specific province names/locations could facilitate identification even in provinces with many households due to the relative granularity of the geographic information that province represents. We decided not to drop province because being able to anonymously distinguish between provinces may be useful for province-level averages and/or regression covariates.	
	~~- (i.e. State)~~ Commune	Avg. pop size (household level): 11	<u>De-identify</u>: There are 14 unique communes represented in the data of which nine have fewer than 10 households which could be identifying. Instead of combining the communes with fewer than 10 households, we decided to randomize all commune IDs because knowing specific commune names/locations could facilitate identification even in communes with many households due to the granularity of the geographic information that commune represents. We decided not to drop commune because being able to anonymously distinguish between communes may be useful for commune-level averages and/or regression covariates.	
	~~- (i.e. Village)~~ Village (lowest)	Avg. pop size (household level): 4	<u>De-identify</u>: There are 35 distinct villages represented in the data of which 31 have fewer than 10 households. Instead of combining the villages with fewer than 10 households, we decided to randomize village ID because knowing specific	

			village names/locations could facilitate identification even in villages with many households due to the granularity of the geographic information that village represents. We do not drop village as being able to anonymously distinguish between villages may be useful for village-level averages and/or regression covariates.	
	~~Lowest - (i.e. Census Blocks)~~ NA	Avg. pop size:	NA	
Knowledge of Treatment	High risk Medium risk Low risk	<u>Low risk</u>: The barymetric survey data contain a treatment variable representing the project's farmer training treatment and comparison zones. Roughly 60 percent of the sample (N=95) is from the project's treatment zone, and the other 40 percent (N=58) is from the project's comparison zone. Given the breakdown and sample sizes, the risk of re-identification using the treatment variable is low, especially considering the comprehensive de-identification of the barymetric restricted-use data.		*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.*
Publication Type	Public-use only Restricted-use only Both	<u>Restricted-use only</u>: Barymetric baseline and one-year follow-up survey data <u>Public-use only</u>: Barymetric key indicators For the justification, please refer to Mathematica's approved data delivery proposal submitted to MCC in <<Burkina public use memo to MCC_revised_final_updated June 27 2019.docx>>, which is included as an attachment to this data delivery package.		*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.*

Section 3: Data Component Preparation Details- Barymetric survey 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]	Financial and agricultural services; Local/regional farmers not selected to participate in farmer training		
2.	What is the potential value to these intruders?	*List all uses (for example:* capture delinquent tax payments, or stigmatize the respondent)	Targeted advertising/cold-call sales; Harassment from farmers not participating in the farmer training program		
3.	What is the expected cost to these intruders to re-identify the data?	*Describe degree of difficulty for re-identification*	It would require more than a significant amount of effort and time for an individual or organization to successfully identify the households in the data, and is therefore unlikely.		

[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*
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*	Farmer training household survey data	*Describe how to mitigate link to existing data that enables re-identification*	The households administered the barymetric baseline and follow-up surveys are a subsample of the households that were administered the farmer training baseline household survey. Although the farmer training baseline household survey data will also be available on a restricted-use basis and can be combined with the barymetric data using the randomized farmer training household ID, the farmer training baseline household data have also been de-identified to prevent re-identification of households and individuals.
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.)	Household ID; Names of respondent and head of household	*List all DIRECT identifiers removed from the dataset.*	Randomized: Household ID Removed: Names of respondent and head of household
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.*	None.	*Describe process for de-identification. For example: introduce random errors into geographic data (GPS, GIS, etc.).*	NA

Specific Issues		Risk Analysis		Risk Mitigation	
		Instructions	*Response*	*Instructions*	*Response*
				Displace urban points 0-2 km, rural points 0-5 km, and 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*	<u>Barymetric restricted-use files</u>: We have prepared the Barymetric survey data for submission to ICPSR as restricted-use files where usage is limited to those with an IRB in place and academic purpose. As such, assessing and masking outlying values as potential indirect identifiers is not necessary (i.e. risk of re-identification via indirect identifiers is mitigated by ICPSR's high restrictions on usage) and would reduce the usability of the data. <u>Barymetric key indicator public-use files</u>: None. Although some of the key indicators in the barymetric public-use files have values greater than three standard deviations from their means, the indicators—cattle age,	*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]*	NA
				Describe any variables that require collapse and describe construction of new variable	None.
				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

[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*
			barymetric measurements, and milk production—are neither sensitive nor potentially identifying at outlying values. No action needed.		
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:*	Barymetric restricted-use files: We have prepared the Barymetric survey data for submission to ICPSR as restricted-use files where usage is limited to those with an IRB in place and academic purpose. As such, assessing and masking unique values or combinations of values is not necessary (i.e. risk of re-identification via unique values or combinations of values is mitigated by ICPSR's high restrictions on usage) and would reduce the usability of the data. Barymetric key indicator public-use files: None. Although some unique combinations of key indicator values exist, the indicators and the unique combinations of their values are neither sensitive	*For each identified rare data, describe the local suppression techniques employed to remove unique and rare data. Specify: are values set to missing, the median, or other?*	NA

Specific Issues		Risk Analysis		Risk Mitigation	
		Instructions	*Response*	*Instructions*	*Response*
			nor potentially identifying. For example, there are only two cows aged 14 in region Boucle du Mouhoun. This unique combination of values does not facilitate re-identification, especially considering the data were collected over five years ago (these cows may have died or been sold/traded since then).		

BURKINA FASO: EVALUATION OF THE BURKINA FASO AGRICULTURE DEVELOPMENT PROJECT
MATHEMATICA
Prepared on: July 12, 2019

Section 2: Data Component Preparation Overview- Farmer training baseline household survey data

	Response		Discussion/Explanation
Data + Code Completeness	Complete	Complete	*To be considered Complete: 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	Baseline only (Separate submission for Interim data, which can be linked using the randomized farmer training household ID)	*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.*
	Interim only		
	Endline only		
	Combination of rounds		

				If combination, please discuss if this file replaces any previously published datasets.
Informed Consent and IRB	High restriction Medium restriction Low restriction	<u>Medium restriction</u>. We received IRB guidance that data can be submitted to ICPSR, and any institution that meets strict ICPSR requirements can access restricted-use data. IRB recommends restricted-use access with at least a secure download restriction.		*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 efforts. Please include any additional guidance provided by the IRB as applicable.*
Geographic Identifiers	Highest (~~i.e. Province~~) Region	Population size (household level): 2164	<u>Identify</u>: There are ten (10) regions represented in the data. De-identifying region would reduce the usability of the data given that region is the highest geographic level represented in the data and thus the most useful geographic identifier for analyses by specific areas of the country for which knowing specific region names would be important. However, we decided to combine contiguous regions with few households—Centre (10 households) combined with Plateau Central (113), Est (2 households) with Centre Est (30), and Nord (14 households) with Centre Nord (303)—resulting in seven (7) unique region categories in the restricted-use farmer training household baseline	*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.*

			survey data. As such, we identify region in the data but combine contiguous non-treatment regions with few households to reduce the likelihood of re-identification.	
	~~- (i.e. District)~~ Province	Avg. pop size (household level): 114	De-identify: There are 19 unique provinces represented in the data of which two have fewer than 10 households which could be identifying. Instead of combining the provinces with fewer than 10 households, we decided to randomize all province IDs because knowing specific province names/locations could facilitate identification even in provinces with many households due to the relative granularity of the geographic information that province represents. We decided not to drop province because being able to anonymously distinguish between provinces may be useful for province-level averages and/or regression covariates.	
	~~- (i.e. State)~~ Commune	Avg. pop size (household level): 58	De-identify: There are 37 unique communes represented in the data of which six have fewer than 10 households which could be identifying. Instead of combining the communes with fewer than 10 households, we decided to randomize all commune IDs because knowing specific commune names/locations could facilitate identification even in communes with many households due to the granularity of the geographic information that commune represents. We decided not to drop commune because being able to anonymously distinguish between communes may be useful for commune-level averages and/or regression covariates.	
	~~- (i.e. Village)~~ Village (lowest)	Avg. pop size	De-identify: There are 88 distinct villages represented in the data of which 25 have fewer than 10 households. Instead of combining the villages	

		(household level): 25	with fewer than 10 households, we decided to randomize village ID because knowing specific village names/locations could facilitate identification even in villages with many households due to the granularity of the geographic information that village represents. We do not drop village as being able to anonymously distinguish between villages may be useful for village-level averages and/or regression covariates.	
	~~Lowest – (i.e. Census Blocks)~~ NA	Avg. pop size:	NA	
Knowledge of Treatment	High risk Medium risk Low risk	<u>Medium risk</u>: The farmer training baseline household survey data contain a treatment variable representing the project's farmer training treatment and comparison zones. Given the original matched-comparison group design, 50 percent of the sampled households are in each zone. The risk of re-identification is generally low given the large sample size in each region (N=1082), especially considering the comprehensive de-identification of the farmer training baseline household restricted-use data. However, for the farmer training households in the Comoé Basin, identification is relatively easier as all 9 villages in the basin were among the treatment villages in the Cascades region (and only those 9). In these communities, the breadth of information from households and community questionnaires would allow parties to identify which of the 9 villages a household is from. Within a village, certain sampled households may be more likely to be identified through family composition, land size, assets or a combination thereof.		*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.*
Publication Type	Public-use only Restricted-use only	<u>Restricted-use only</u>: Farmer training baseline household survey data <u>Public-use only</u>: Farmer training household key indicators		*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.*

	Both	For the justification, please refer to Mathematica's approved data delivery proposal submitted to MCC in <<Burkina public use memo to MCC_revised_final_updated June 27 2019.docx>>, which is included as an attachment to this data delivery package.	

Section 3: Data Component Preparation Details- Farmer training baseline household survey 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]	Government officials; Financial and agricultural services; Local/regional farmers not selected to participate in farmer training		
2.	What is the potential value to these intruders?	*List all uses (for example:* capture delinquent tax payments, or stigmatize the respondent)	Tax payments; Targeted advertising/cold-call sales; Harassment from farmers not participating in the farmer training program		
3.	What is the expected cost to these intruders to re-identify the data?	*Describe degree of difficulty for re-identification*	It would require a significant amount of effort and time for an individual or organization to successfully identify the households in the data.		

[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*
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*	Farmer training supplemental household survey data; Barymetric baseline and one-year follow-up survey data	*Describe how to mitigate link to existing data that enables re-identification*	Disparate subsamples of the households administered the farmer training baseline household survey were administered the farmer training supplemental household survey and the barymetric baseline and one-year follow-up surveys. Although the data of those surveys will also be available on a restricted-use basis and can be combined with the farmer training baseline household survey data using the randomized farmer training household ID, those data have also been de-identified to prevent re-identification of households and individuals.
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.)	Household ID; Names of respondent, head of household, household members, plots, and banks; location of household member (if not currently in household); Any "other, specify" variables with text responses containing PII; Data collector names and IDs	*List all DIRECT identifiers removed from the dataset.*	Randomized: Household ID Removed: Names of respondent, head of household, household members, plots, and banks; location of household member (if not currently in household); Any "other, specify" variables with text

Specific Issues		Risk Analysis		Risk Mitigation	
		Instructions	*Response*	*Instructions*	*Response*
					responses containing PII; Data collector names and IDs
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.*	Latitude and longitude	*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[2].*	We dropped latitude and longitude to prevent re-identification.
7.	**Attribute Disclosures:** What variables have OUTLIERS that create INDIRECT identifiers are in the raw data?	*List the identifying items/variables*	<u>Farmer training baseline household restricted-use files</u>: We have prepared the farmer training baseline household survey data for submission to ICPSR as restricted-use files where usage is limited to those with an IRB in place and academic purpose. As such, assessing and masking	*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]*	<u>Farmer training baseline household key indicator public-use file</u>: Yields per hectare and agricultural profit were top-coded at three standard deviations above the mean. Agricultural profit was also bottom-coded at three standard deviations below the mean.

[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*
			outlying values as potential indirect identifiers is not necessary (i.e. risk of re-identification via indirect identifiers is mitigated by ICPSR's high restrictions on usage) and would reduce the usability of the data. <u>Farmer training baseline household key indicator public-use file</u>: Outlying values of the key indicators that are continuous variables—yields (per ha) and agricultural profit—could potentially be indirect identifiers.	*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.	NA
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:*	<u>Farmer training baseline household restricted-use files:</u> We have prepared the farmer training household survey data for submission to ICPSR as restricted-use files where usage is limited to those with an IRB in place and academic purpose. As such, assessing and masking unique values or combinations of values is not necessary (i.e. risk of	*For each identified rare data, describe the local suppression techniques employed to remove unique and rare data. Specify: are values set to missing, the median, or other?*	<u>Farmer training baseline household key indicator public-use file:</u> Rare crops: Only three households grew cowpeas in the dry season. We replaced the cowpea yields per hectare with missing values for these three households in that season to prevent potential re-identification.

Specific Issues		Risk Analysis		Risk Mitigation	
		Instructions	*Response*	*Instructions*	*Response*
			re-identification via unique values or combinations of values is mitigated by ICPSR's high restrictions on usage) and would reduce the usability of the data. <u>Farmer training baseline household key indicator public-use file:</u> Specific crops grown and irrigation types used by few households could potentially be indirect identifiers.		Rare irrigation types: Mobile boom and pivot irrigation were used by few households in both the dry and rainy seasons. We combined those irrigation types with "other" irrigation to prevent potential re-identification.