NAMIBIA: COMMUNITY SKILLS DEVELOPMENT CENTERS (COSDEC) EVALUATION
MATHEMATICA POLICY RESEARCH
Prepared on: 09/15/2017

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*.)

This data package pertains to data collected from individuals who enrolled in national courses starting between July and December 2014 in the seven Community Skills Development Centers (COSDECs) that benefitted from the MCC's COSDEC subactivity in Namibia. These individuals were surveyed about one year after the end of training. Survey Warehouse, a local data collection firm, conducted the survey from January to June 2016. The survey collected data on enrollees' demographic characteristics, as well as their vocational training history, employment status, and earnings and income.

The data package includes the following components:

1. Public use data file
2. Restricted use data file
3. Cleaning do file
4. Construct and analysis do file
5. Codebook summary and full codebook
6. Users' manual
7. Training manual
8. Metadata (Nesstar file)
9. Questionnaires
10. Transparency statement

Note: The data package does not include the raw data, which contains personally identifiable information such as names and national identification numbers, because this information is not necessary to replicate our 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*.)

This data package considers the following complementary data efforts:
None.

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**	**Other Documents**
2016 Public Use Data Package	COSDEC_followup_DRBCover.docx	Pg 1 of questionnaire	COSDEC_followup_ ENGLISH	• COSDEC_cleaning.do (cleaning do file) • COSDEC_analysis.do (constructs and analysis do file) • COSDEC_followup_PUF_codebook.txt (codebook) • COSDEC_followup_PUF_codebook_ contents.xlsx (codebook summary) • COSDEC_followup_users_manual.pdf (users' manual) • COSDEC_followup_training_manual.doc (training manual) • COSDEC_followup.Nesstar (metadata) • COSDEC_followup_transparency_ statement.pdf (transparency statement)

Section 2: Data Component Preparation Overview

	Response		**Discussion/Explanation**
Data + Code Completeness	Complete	**Complete.** We have provided the full dataset (including constructed variables) and analysis code. The anonymization procedures did not affect constructed variables for our key primary and secondary outcomes. Users should therefore be able to exactly replicate the key findings in our report.	*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.*
	Incomplete		*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.*
Data Round(s):	Baseline only	**Endline only.** Only one round of the survey was conducted, as this was sufficient to inform the planned outcomes analysis. No additional rounds are planned in the future.	*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.** The consent statement states that *"Any information you provide that can identify you will be kept strictly confidential by the parties conducting this study […] These users will use data for statistical purposes only. Once the study is completed, all data from the study that does not identify you personally will be made publicly available for others to study."* This language suggests that MCC can make the data	*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		

		publicly available as long as there are no direct identifiers. The de-identification efforts we describe here are designed to minimize the reasonable risk of identification, even with indirect identifiers.		
Geographic Identifiers	Highest (i.e. Province)	Region, pop 70,000-350,000	**Identify.** Region of origin poses a minimal risk to identification given the number of observations per region (a minimum of 14 after combining two regions with fewer than 10 observations into the "other" category) and the large population size of the regions.	*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.*
	--(i.e. District)	NA	NA	
	--(i.e. State)	NA	NA	
	--(i.e. Village)	Unknown	**Remove.** Town of origin often has no more than a handful of observations per value and therefore poses a risk to identification. Collapsing values into an "other" category would eliminate much of the variation in this variable, and would be of little value to users of the data.	
	Lowest --(i.e. Census Blocks)	NA	NA	

Knowledge of Treatment	High risk	**NA.** There was no counterfactual and all respondents were treated.	*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	**Both.** The pubic use file includes all the information necessary to replicate our key findings, while taking additional de-identification efforts to protect the confidentiality of respondents. The restricted use file includes: (1) additional information regarding training (course name and timing), and (2) all variables removed or adjusted in the public use file, except for direct identifiers. This information can be merged to the public use file through a unique identifier for each respondent. Users with specific data needs might find the additional information in the restricted use file valuable. For example, users would be able to examine outcomes for specific courses.	*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		

Section 3: Data Component Preparation Details

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]	There are no significant incentives to de-anonymize survey respondents.		
2.	What is the potential value to these intruders?	*List all uses (for example:* capture delinquent tax payments, or stigmatize the respondent)	The re-identified data would have limited value. It is possible that they could be used by the tax authorities to identify tax-evaders. However, wages and income are generally low for this sample, and the data could not be used to identify individuals with very high wages/incomes because we have top-coded this information in the public use file. Therefore, the incentive to identify tax-evaders would be limited.		

[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*	It would require considerable effort on the part of an intruder to identify respondents based on certain combinations of responses. The financial, legal, and cultural cost to respondents, if identified, would be low.		
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*	Because some courses only had a small number of enrollees, including detailed training information such as course name and timing would potentially allow respondents to be identified in combination with information on their individual characteristics and outcomes.	*Describe how to mitigate link to existing data that enables re-identification*	We removed the information on course name and replaced with an arbitrary code (information on training provider does not pose a risk because the number of enrollees per provider is large). We also removed the course start and end dates, but retained a construct for course duration. The detailed training information is still available to approved users in the restricted use file.
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.)	1. Individual identifiers: name and national ID number. 2. Geographic identifiers: region and town of origin	*List all DIRECT identifiers removed from the dataset.*	1. All individual identifiers were removed from both the restricted and public use files. Date of birth of the respondent and their children (if any) were also removed from the public use file as an additional precaution because they are unique in many cases

Specific Issues		Risk Analysis		Risk Mitigation	
		Instructions	*Response*	*Instructions*	*Response*
					and could be used as identifiers. 2. Data were collected from training applicants who are located throughout Namibia; the sample is therefore not nested in geographic units. Because the town of origin has only a handful of observations in many cases, we have removed it from the public use file. Retaining the region of origin poses minimal identification risks to individual respondents, because these geographic units have a relatively large population size and there is a large number of observations per region (a minimum of 14 after combining two regions with fewer than 10 observations into the “other” category). Therefore, because the risk of identification is minimal and this variable could be useful for research purposes (for example, in conducting subgroup analyses by regional

Specific Issues		**Risk Analysis**		**Risk Mitigation**	
		Instructions	*Response*	*Instructions*	*Response*
					characteristics), we chose to retain it in the public use file.
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.).* *Displace urban points 0-2 km, rural points 0-5 km, and additional 1% of rural points 0-10 km[2].*	NA
7.	**Attribute Disclosures:** What variables have OUTLIERS that create INDIRECT identifiers are in the raw data?	*List the identifying items/variables*	To objectively and systematically identify variables that posed a risk to identification, we identified all continuous variables as well as all discrete variables for which a given value was reported by fewer than 10 respondents. Then, on a case by case basis, we determined whether and how these variables should be adjusted to limit the risk of direct identification.	*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]*	1. Household size: because large households were uncommon, we collapsed these into categories for 13-15 members, 16-19 members, and 20 or more members. 2. Age: upper and lower bounds were set as the 95th and 5th percentiles for the full sample, respectively. 11. Wages, earnings from self-employment, other individual income, other household income: upper bounds for each variable were set as the

[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*
			The variables that we adjusted are as follows: 1. Household size 2. Age 3. Level of education 4. Vocational training provider 5. Duration of training 6. Skill area of training 7. Level of training 8. Level of summative assessment 9. Duration of use of small/medium enterprise (SME) unit 10. Type of work 11. Wages, earnings from self-employment, other individual income, other household income 12. Number of dependents 13. Marital status 14. Region of origin 15. Language group We also removed all “other, specify” text responses, as many of these were specific to a small number of respondents. The responses are also difficult to understand. With		95th percentile for the full sample. 12. Number of dependents: because a large number of dependents was uncommon, we collapsed these into categories for 8 or 9 dependents, 10-14 dependents, and 15 or more dependents.
				Describe any variables that require collapse and describe construction of new variable	3. Level of education: levels below grade 8 completion were rare, and we collapsed all of these into a single category. Higher education (above grade 12) was also rare, and we collapsed this with the grade 12 category. 4. Vocational training provider: because few respondents enrolled in additional trainings provided by non-COSDEC providers, these providers were collapsed into a single “other” category. 5. Duration of training: self-reported durations 18 months or longer were rare and were collapsed into a single

Specific Issues	**Risk Analysis**		**Risk Mitigation**	
	Instructions	*Response*	*Instructions*	*Response*
		substantial effort we could possibly back-code some of these into new categories, but even that would not yield many categories above the cutoff of 10 observations and would be of little value to a user of these data. We determined that the following variables did not require adjustment, despite having less than 10 observations because the responses were unlikely to be public information and/or are relative to a specific period defined by the survey date, which was conducted more than a year ago. They therefore pose a low risk to identification. 16. Number of training programs attended since July 2014 17. Reason for dropping out of training 18. Duration of job attachment		category. For the duration based on administrative data, the 3 month category (which had less than 10 observations) was combined with the 2 month category. 6. Skill area of training: any skill area reported by fewer than 10 respondents (across all trainings) was collapsed into an “other category”. 7. Level of training: few respondents had a training at levels 3, 4, or 5. We combined all of these into one category. 8. Level of summative assessment: few respondents had an assessment at levels 3, 4, or 5. We combined all of these into one category. 9. Duration of use of SME unit: because few respondents used the SME units, the reported duration of use had many small categories. We therefore combined these into two categories: 0-4 weeks and 4 or more weeks.

Specific Issues		**Risk Analysis**		**Risk Mitigation**	
		Instructions	*Response*	*Instructions*	*Response*
			19. Opinions on training quality (Likert scale) 20. Whether passed a summative training assessment 21. Details of SME unit use (reasons for use, perceived benefits, support provided, etc.) 22. Plans for additional vocational training in the next 2 years 23. Number of jobs held at the survey date and in the 3 years before the survey 24. Start and end dates of jobs in the 3 years before the survey 25. How learned about job 26. Whether paid for work 27. Hours per week worked 28. Reason for not being available for work/not working 29. Job tenure 30. Time required to find job since end of training		10. Type of work: any type of work reported by fewer than 10 respondents (across all jobs) was collapsed into an “other category”. 13. Marital status: rare marital status categories (e.g. divorced) were combined into broader categories. 14. Region of origin: two regions were reported by fewer than 10 respondents and were collapsed into the “other” category. 15. Language group: we collapsed any languages reported by less than 10 respondents into the “other” category.
				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.	All recoding is described in the row above.
8.	**Attribute Disclosures:** What variable combinations produce UNIQUE observations that create INDIRECT IDENTIFIERS (for example: individuals with high	*List the identifying items/variables:*	With our cutoff of a minimum of 10 observations per value for the variables identified above, the risk of a two-	*For each identified rare data, describe the local suppression techniques employed to mitigate the identification risk of unique*	Given the limited risk to respondent identification, we do not recommend further exploration of potential privacy risks by conducting

Specific Issues		**Risk Analysis**		**Risk Mitigation**	
		Instructions	*Response*	*Instructions*	*Response*
	incomes, ages, or unique combinations, such as 17-year old widowers or contextually unusual racial/ethnic backgrounds)		way cross tabulation being used to identify unique respondents is low. In addition, exploring these cross-tabulations in detail would require a high level of effort on our part (and even more effort would be required to explore higher-level cross-tabulations, like unique combinations of three variables). Because this would also require a high level of effort from an intruder—with no obvious incentives to do so—we believe that the risk to respondents of being identified in this manner is very low in practice.	*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.)*	cross-tabulations in this instance.

[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.