MCC Evaluation Microdata
Data Package

Section 1: Cover Sheet

Overview of Data Package

Note:
For each component, sections 2 and 3 are attached as separate word documents.

As described in more detail in the individual worksheets, the public data sets have been substantially altered to minimize the risk of re-identification of TVET teachers and school staff. Geographic identifiers have been removed and school names anonymized. The trades offered by schools have been recoded into broader categories and outliers in continuous variables recoded to the 1st or 99th percentile or both, where applicable. Binary and discrete variables with a small number of observations per category have been recoded or removed. Most string variables have been dropped from the publicly available data.

All student level data sets can be uniquely merged through the variable *id_new*.
All administrative data sets can be merged through the variable *schoolcode*.

The components of this data package were collected according to the following schedule

Cohort	Program	Year of Data Collection					
					Wave 2	Wave 3	Wave 4
		2010	**2011**	**2012**	**2013**	**2014**	**2015**
2010	1 year	Admissions			GFU	Tracking	Tracking
	2/2.5 year	Admissions			GFU	Tracking	Tracking
2011	1 year		Admissions		GFU	Tracking	Tracking
	2/2.5 year		Admissions		Tracking	GFU	Tracking
2012	1 year			Admissions	Tracking	GFU	Tracking
	2/2.5 year			Admissions	Tracking	Tracking	GFU
Administrative						Secondary Management Teacher	

The publicly available data package includes the following components:

1. Teacher survey wave 3

Individual level data: The teacher survey was administered to TVET school teachers and includes data on their socio-economic characteristics, impressions of classroom, equipment and curriculum quality, student ability and behavior and perceptions of the TVET reforms implemented by MCA-M. Wave 3 was administered in 2013.

2. Secondary survey wave 3

School level data: The Secondary Survey captured information on enrollment and graduation numbers, curricula, private-public partnerships and grants, equipment, funding sources and other donor activities. The questionnaire was answered by administrative TVET school staff. Data collection Wave 3, 2013.

Note: Merges with other data sets through the variable *schoolcode*

3. Management surveys wave 3

Individual level data: The Management survey was administered to senior school management staff. It includes data on staff's socio-economic characteristics and information on tuition costs, enrollment and graduation rates, student to teacher ratios, perceived teacher competence and availability and use of training equipment. Data collection Wave 3, 2013

Note: Merges with other data through the variable *schoolcode*

The restricted data package contains the following data sets:

1. Admissions surveys 2010, 2011 and 2012

Individual level data: the Admissions questionnaire recorded students' social, economic and demographic characteristics and served as this evaluation's baseline data set. It was administered to applicants to this study's 10 evaluation schools in the years 2010, 2011 and 2012. The questionnaire included the application form for the 10 schools that participated in this study's admissions lotteries, which was removed from the publicly available data.
Key variables the data includes: applicants' anonymized ID variables, identical to ID variables in the graduate follow up and tracking surveys; treatment indicator and probability of treatment; lottery year, round and anonymized school code. It is the only data set that includes students' treatment status (*TVET_accepted*), which was determined by admissions lotteries at all school in each year and round of admissions. The probability each student had of being admitted to a school they applied to through the lottery is given by *TVET_accepted_p*. *assigned_trade* denotes the (recoded) trade students admitted through the lottery were assigned.

Note: the variables *schoolcode*, *year* and *round* indicate the year and round respondents applied to a (anonymized) school. The variable *schoolcode* merges with all other data sets in this data package.

2. Graduate follow up (GFU) surveys waves 2, 3 and 4

Individual level data: as the primary instrument to capture post-graduation labor market outcomes, the GFU survey was administered in person about one year after each cohort's theoretical graduation date to learn about the short-term impacts of studying improved trades. It includes detailed information on students' education, employment, income, assets, monetary and in-kind transfers and results of brief trade-specific tests they took.
Wave 2 was administered in 2013; Wave 3: 2014; Wave 4: 2015.

Note: The Graduate follow up survey merges with the Admissions survey through i*d_new*.

3. Tracking surveys waves 2, 3, 4

Individual level data: The Tracking survey primarily collected respondents' latest contact information and that of their parents, relatives and friends (all of which was removed from the public data).

It also included questions about students' education, employment and income. The tracking survey was administered over the phone. Wave 2 in 2013; Wave 3 in 2014; and Wave 4 in 2015.

Note: The Tracking survey merges with the Admissions survey through the variable "id_new". Variable prefix *t4k* denotes Wave 4 data; *t3k* wave 3; *t2k* wave 2

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:

- Complementary Component 1:
- Complementary Component 2:

Data Package Folder Contents

(*Instructions: Please list the 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 1: Data Package Components

Data Package			
Component	**Worksheet**	**Informed Consent and Questionnaire**	**Other Docs**
1_DRB_Teacher	Teacher survey	Teacher questionnaire	
2_DRB_Management	Management survey	Management questionnaire	
3_DRB_Secondary_information	Secondary information survey	Secondary information questionnaire	
4_DRB_Admissions-Restricted_use	Admissions surveys	Admissions questionnaire	
5_DRB_GFU-Restricted_use	Graduate follow up surveys	Graduate follow up questionnaire	
6_DRB_Tracking-Restricted_use	Tracking surveys	Tracking questionnaire	

Section 2: Secondary Information Survey Preparation Overview

	Response		Discussion/Explanation
Data + Code Completeness	Complete	Incomplete: some variables were recoded to reduce the possibility of re-reidentification through outliers or variables with few observations. The data set used for analysis contained the full list of 66 trades on offer at TVET schools. In the public data set, trade variables are recoded into seven broader categories to decrease the likelihood of re-identification.	*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	Wave 3 data: Secondary Information survey administered in 2013	*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	Medium restriction. In addition to all direct identifiers, we removed a large number of indirect	*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*

	Medium restriction	identifiers to protect the right to privacy of respondents as outlined by our informed consent.		*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	Highest: TVET school	Avg. pop size: 50-200 staff	De-identify school Identifying school would increase the likelihood of staff re-identification	*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)	Avg. pop size	N/A	
	--(i.e. State)	Avg. pop size	N/A	
	--(i.e. Village)	Avg. pop size	N/A	
	Lowest --(i.e. Census Blocks)	Avg. pop size	N/A	
Knowledge of Treatment	High risk	N/A (Secondary survey does not include treatment information)		*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: anonymized public use data set and restricted data set used for analysis by contractor and qualified researchers		*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: Secondary Survey 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]	The re-identification risk is relatively low, principally because the value of re-identification to an outside party is low. School administrators working at the time of the interview are those most likely to be able to recognize and re-identify their own schools, but school administrators were unlikely to learn anything they did not already know. Even more practically, the last survey was in 2013 and many school administrators have since moved on.		
2.	What is the potential value to these intruders?	*List all uses (for example:* capture delinquent tax payments, or stigmatize the respondent)	NA		
3.	What is the expected cost to these intruders to re-identify the data?	*Describe degree of difficulty for re-identification*	Medium Current or former school administrators with		

[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*
			detailed knowledge of their schools' characteristics could re-identify their own schools. They would, however, be unlikely to learn anything they didn't already know.		
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*	The Secondary Survey data can be linked to other TVET data sets via anonymized school codes. The difficultly of re-identifying remains high. Re-identification of schools is difficult, and further linking of staff within these schools would be even more challenging due to a lack of publicly available data sets.	*Describe how to mitigate link to existing data that enables re-identification*	All direct and indirect identifiers removed from data collected by the contractor
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.)	Names; contact details; addresses; phone numbers; government-issued ID numbers;	*List all DIRECT identifiers removed from the dataset.*	Names; addresses; contact details; phone numbers; government-issued ID numbers;
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	*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].*	N/A

[2] 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*	N/A. The Secondary survey did not include integer data about personal characteristics.	*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]	
				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 widowers or contextually unusual racial/ethnic backgrounds)	*List the identifying items/variables:*	Position, work experience, trade information	*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?*	Removed position, work experience Recoded the trades offered by the schools into broader categories to increase difficulty of re-identifying of schools

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

MCC Mongolia TVET Equipment Upgrade
QA of De-Identification and Review of Data
Joseph Green, Independent Data Consultant
Prepared on: December 20, 2019

MCC Impact Evaluation Microdata
Data Package

Summary of Previous Virtual Reviews

This data package has been through several iterations of preparation and review for publication. Jen Sturdy summarized the results of the past virtual reviews:

> This has been a complicated review given it is a large dataset at multiple levels – school, administrator, teacher, student – with multiple surveys at some levels, all of which can be linked across surveys, and with previously posted de-identified baseline data.
>
> In summary, the M&E Technical Team raised several issues with the IPA evaluation team (i) prior to the 6/23/2017 DRB (where this package was discussed but there was no quorum) and (ii) following the 10/13/2017 DRB (where M&E removed this package from the agenda given incomplete responses to previous issues raised). The main issues were:
>
> 1. For school, administrator, teacher data – Can the school name be de-identified? Is the data included in these surveys sufficiently de-identified to avoid risk to respondents?
>   - IPA Response: This refers to these three datasets in the package: Teachers; Management; Secondary information.
>
>   After review of EXISTING, AVAILABLE data, IPA determined that school names **cannot be de-identified** given information on school website's admissions data and the already published baseline report. However, to reduce re-identification risk for teachers and administrators, trades-taught and other identifiers (listed in Worksheet) have been removed or recoded to significantly reduce re-identification risk for individuals.
>
> 2. For student data – Would it be possible for "insiders" to re-identify students based on known trades? Other linkage points (such as knowledge to treatment)?
>   - IPA Response:
>   This refers to these three datasets in the package: Admissions; Graduate Follow-Up (GFU); and Tracking (TRK).
>
>   - **Graduate Trades**. IPA suggests grouping Trades into broader categories. Rather than including the specific trade names, they can group trades into broader categories. They used broader categories for analysis of the impact of upgraded equipment (re-using these isn't a good option as individual trades are often present in multiple categories). The advantage is clear: it would be even more difficult for persons with prior knowledge of respondents (schoolmates or school administrators) to link observations with specific individuals. Grouping would prevent the full replication of the results in the report using individual trades, but reduces re-identification risk with school name.

- **Linkage docs and re-identification risk**. The student IDs in the website's admissions data are different from the newly prepared student IDs. However, since the admissions data set is quite rich, finding a way to link the two is probably feasible (once merged with the newly anonymized admissions data, it could then easily be merged with the graduate follow up and tracking data). It is significantly harder to merge the website's admissions data directly with the graduate follow up and tracking surveys. Not uploading an updated admissions data set that links up with the follow up surveys would reduce risk, at the cost of depriving users of working with baseline and follow up data jointly.
- **Result**: Re-identification risk is considered low, but possible, therefore they suggest producing restricted-access ONLY student data which retains all necessary information and manages risk through restricted-access.

Summary of QA and Review of Data

This data package includes the following components:

- 1. Admissions survey: 1-TVET-Admissions-DRB.dta
- 2. Graduate follow-up survey: 2-TVET-GFU-DRB.dta
- 3. Tracking survey: 3-TVET-TRK-DRB.dta
- 4. Secondary information survey: 4-TVET-Secondary-DRB.dta
- 5. Management survey: 5-TVET-Management-DRB.dta
- 6. Teacher survey: 6-TVET-Teacher-DRB.dta

My task was to evaluate the data for publication given the context of the past reviews. One re-identification risk to determine for all files is whether or not insiders or outsiders could re-identify the respondents (or schools) of each survey. To tackle this question, I identified the variables that insiders (respondents) and outsiders (anyone else) could know about the respondents, and in turn use to identify the respondents. For example, in the Teacher data, some of the variables whose values both insiders and outsiders could know for particular respondents include age, gender, bachelor's degree (yes/no), master's degree, type of trade taught (8 categories), and a few other variables. In addition, some variables only insiders could know are authority to teach (which I assumed is a certification), funding source, and training seminar topics. The specific variables tagged for each survey are detailed in the "36 – Mongolia TVET – data info.xlsx" file.

I then determined what percentage of the survey's respondents had unique values across all of these insider or outsider varaibles. If, for example, a high percentage of respondents in a survey had unique values across all of the outsider varaibles, then that survey's data poses a high risk of re-identification by outsiders. The results of this analysis are detailed in "Attachment A – Data Uniqueness.pdf" file. In short, all surveys except for the Tracking survey have a high percentage of unique respondents that insiders could possibly identify. The Teacher and Management data also have a high percentage of unique respondents across the variables that outsiders could identify.

Unfortunately most of these variables are also very useful for analysis, so removing or altering them beyond what was done by IPA to reduce their re-identification risk would severly limit the usability of the data.

I summarize my findings for each component data set below.

1. Admissions survey: 1-TVET-Admissions-DRB.dta

IPA stated that the Admissions survey data had many linkable variables to the school's website admissions data, and would be very difficult to fully de-identify. My data uniqueness analysis supports this conclusion, showing that 100% of respondents are unique across the insiders variables, and 11% of respondents are unique or nearly unique (groups of 4 or fewer) across the outsiders variables. The informed consent states, "Any information that can identify you (or your legal ward) individually will be kept strictly confidential. It will only be known to those conducting the study", I do not believe we should release this data publicly. Given that all information that can identify individuals would be very difficult to remove from the data, the DRB should consider whether those with restricted-access qualify as "those conducting the study."
Conclusion: I agree with IPA's suggestion to produce a restricted-access ONLY admissions data which retains all necessary information and manages risk through restricted-access.

2. Graduate follow-up survey: 2-TVET-GFU-DRB.dta

Based on my unique respondent analysis, this data appears to be high risk for re-identification by insiders. 93% of repondents are unique or nearly unique (groups of 4 or fewer) for insiders. However, there is less risk of re-identification by outsiders, as only 3 of the nearly 11,000 respondents are nearly unique. If the DRB decides to produce a restricted-access version of this data, I do not think it has any sensitive variables that need to be altered.
Conclusion: I suggest a restricted-access ONLY graduate follow-up data which retains all necessary information and manages risk through restricted-access.

3. Tracking survey: 3-TVET-TRK-DRB.dta

Based on my unique respondent analysis, this data appears to be very low risk for re-identification by either the insider or outsider variables. However, I trust IPA's assessment that the complementary Admission's website data could be linked to this data, and that this data should only be released with restricted access. If the DRB decides to produce a restricted-access version of this data, I do not think it has any sensitive variables that need to be altered.
Conclusion: I suggest a restricted-access ONLY graduate follow-up data which retains all necessary information and manages risk through restricted-access.

4. Secondary information survey: 4-TVET-Secondary-DRB.dta

This school-level data has many variables that could identify the schools, but none that are sensitive about the schools themselves. However, the informed consent does not promise that the school's identity will remain undisclosed. The only issue I can see with releasing this data is that if any of the other files

were also released publicly, the identify of those schools could be revealed by linking to this data through the 'anonymized' schoolcode variable (and to a lesser extent, other variables they have in common). I do not recommend that any of the other files be released publicly, but in case the DRB decides to release any of them, we can at least prevent the schools' identities revealed in this Secondary data from being linked to the other data by removing or masking the linkage variables.
Conclusion: I suggest releasing this data publicly without any additional changes.

5. Management survey: 5-TVET-Management-DRB.dta

This individual-level data of senior school management staff is high-risk for re-identification for insiders and outsiders based on my unique respondent analysis. 95% of respondents could be uniquely identified by insiders, and 58% of respondents could be uniquely or nearly uniquely (groups of 4 or fewer) identified by outsiders. I identified (in blue highlight in the "36 - Mongolia - TVET - data info.xlsx" file) a further 23 variables that contain uniquely-identifying open-ended responses, including variables that identify the names of teachers and training organizers. The informed consent for this survey allows the full, identified data to be used by "researchers working with the data". Given that the data would be hard to de-identify sufficiently and still be usable, I suggest this data only be release with restricted access.
Conclusion: I suggest a restricted-access ONLY management data which retains all necessary information and manages risk through restricted-access.

6. Teacher survey: 6-TVET-Teacher-DRB.dta

This individual-level data of senior school management staff is high-risk for re-identification for insiders and outsiders based on my unique respondent analysis. 100% of respondents could be uniquely identified by insiders, and 97% of respondents could be uniquely identified by outsiders. I identified (in blue highlight in the "36 - Mongolia - TVET - data info.xlsx" file) a further 59 variables that contain uniquely-identifying codes or open-ended responses. These include variables that identify the names of equipment categories that are nearly as specific as the original trades IPA aggregated into 8 broad categories. The informed consent for this survey allows the full, identified data to be used by "researchers working with the data". Given that the data would be hard to de-identify sufficiently and still be usable, I suggest this data only be release with restricted access.
Conclusion: I suggest a restricted-access ONLY teacher data which retains all necessary information and manages risk through restricted-access.

Complementary Data

This data package considers the following complementary data efforts:

- 1. School website's admissions data.
- 2. De-identified baseline data, posted previously.
- 3. Mongolia TVET Vocational data package prepared by Social Impact in July 2019, not yet released.

Data Package Folder Contents

Data Package			
Component	**Worksheet**	**Informed Consent and Questionnaire**	**Other Docs**
1-TVET-Admissions-DRB.dta	36 – QA of Mongolia TVET De-Identification.docx	Admissions questionnaire	36 – Mongolia TVET – data info.xlsx Attachment A – Data Uniqueness.pdf
2-TVET-GFU-DRB.dta		Graduate follow up questionnaire	
3-TVET-TRK-DRB.dta		Tracking questionnaire	
4-TVET-Secondary-DRB.dta		Secondary information questionnaire	
5-TVET-Management-DRB.dta		Management questionnaire	
6-TVET-Teacher-DRB.dta		Teaching questionnaire	