Tanzania Education Statistical Disclosure Control Process

External Report

1. Disclosure risk and confidentiality protection

Microdata often contain confidential or sensitive information, which makes release of these datasets in their original form impossible. Release of the data could reveal this confidential information and lead to a breach of privacy of the respondents. This has ethical and in many cases legal objections. Furthermore, when confidentiality in not guaranteed, current and potential future respondents are less likely to be willing to respond in future surveys.

Statistical Disclosure Control (SDC) refers to (i) a set of methods to measure the risk in a dataset that confidential information may be disclosed when releasing a dataset and (ii) to the set of methods to treat the data in order to prevent the release of confidential information when releasing the dataset. SDC is used in many statistical offices to anonymize data before release. Templ et al. (2014) give a concise introduction to SDC. We refer to Hundepool et al. (2012) and the references therein for a complete overview of the SDC process and a detailed description of the methods.

In this report we evaluate the risk of disclosure to the SDI Tanzania education dataset and describe the methods applied to mitigate this risk. Anonymization leads to changing the dataset. Therefore, we evaluate the validity of the dataset after the anonymization by evaluating common (published) SDI indicators, which are derived from the data. The aim of this study is to create a Public Use File (PUF), which is a dataset that is freely accessible to the greater public.

2. SDI Tanzania education data set

The SDI Tanzania education dataset contains information on a sample of 400 schools from a total of about 14,600 schools in 22 regions. The statistical objects in this dataset are schools, teachers and pupils. The main concern for re-identification and confidentiality are the teachers, and, to a lesser degree, pupils. However, since the data are hierarchical, i.e. teachers and pupils belong to schools, the re-identification of a school might lead to the re-identification of a teacher or pupil too.

The datasets consists of the following modules that contain information on:

- Module 1: School information
- Module 2: Teacher roster
- Module 3: Financial information
- Module 4: Classroom observation
- Module 5: Pupil assessment
- Module 6: Teacher assessment

These datasets contain sensitive and confidential variables, especially on the level of the

teacher but also on the school level (expenditures and revenue). Due to the confidentiality and high probability of re-identification, Module 3 is not released as PUF and only available as a Scientific Use File (SUF) for trusted researchers.

3. Data preparation

The SDC process is done with the aid of the statistical software *R*.[1] The *R* package *sdcMicro* with functions for risk measurement and the application of SDC methods has been developed by Templ et al.[2] All anonymization steps are reproducible with the R script for Tanzania. The anonymization process starts from the harmonized dataset. Before the start of the anonymization process, any final data quality corrections are made. For example, missing value codes other than the standard *R* missing value code *NA* are recoded to be correctly interpreted by the software. Table 1 provides an overview of the values by variable that have been recoded. *STATA* missing values are automatically recoded when reading the data in *R*.

Table 1 Missing value recodes

Variable	**Code(s)**
m1saq6 (when did the school begin operating?)	-8
m1sbq6b, m1sbq6c (How many times did the student government meet in 2013?; Are minutes/action plans from the latest student government meeting published in public?)	-9
m1sbq6da, m1sbq6db, m1sbq6dc, m1sbq6dd, m1sbq6dg (In what ways can students at this school express their grievances? different ways)	-3
m1sbq6dh (In what ways can students at this school express their grievances? different ways)	-3, -8, -9
m1sbq8a (Did you receive any written feedback/check list from the quality assurance officer?)	-8
m1scq4a, m1scq4b, m1scq5, m1scq6, m1scq8 (Questions relating to number and state of toilets)	-9
m1scq9 (Questions relating to number and state of toilets)	-9, -99
m1scq12 (What means of transport do you usually use to get to the district education office?)	-9
m2sbq20, m2sbq21, m2sbq22b, m2sbq22c (number of classrooms, different categories)	-9
m2saq7a, m2saq7b, m2saq7c (taught math/English?)	-9
m2sbq4, m2sbq5, m2sbq7, m2sbq8, m2sbq9a, m2sbq9b, m2sbq9c, m2sbq9d (reason for absence, position in school, in which classes do you teach?)	-9
m2sbq12 (in what year did you start teaching?)	-999
m2sbq10, m2sbq11, m2sbq13, m2sbq14 (education level, age, born in district, age)	-9
m2sbq15, m2sbq15a1, m2sbq15a2, m2sbq15a3, m2sbq15a4 (unpaid claims)	-9
m2sbq16, m2sbq17a, m2sbq17b, m2sbq17c, m2sbq17d, m2sbq18 (comments from head teacher/administrator)	-9

[1] Available from https://www.r-project.org/.
[2] The package is available from CRAN (https://cran.r-project.org/).

m4sbq1, m4sbq4, m4sbq4a, m4sbq4b, m4sbq5, m4sbq5a, m4sbq5b, m4sbq6, m4sbq6a, m4sbq6b	-9
m4sbq17	-99999
m4scq1, m4scq, m4scq2, m4scq2a, m4scq2b, m4scq3, m4scq4, m4scq6, m4scq6a, m4scq6b, m4scq7, m4scq8, m4scq9, m4scq11, m4scq13, m4scq14, m4scq15, m4scq16, m4scq17, m4scq18, m4scq19, m4scq20, m4scq21, m4scq22, m4scq23, m4scq24, m4scq25, m4scq26	-9
m4sdq3	-9
m6siq8 , m6siq10a, m6siq10b, m6siq10c, m6siq11a, m6siq11b, m6siq11c	-9

Other checks include reviewing outliers and dealing with facilities that may have refused to complete a part of the questionnaire.

4. Anonymization measures

This dataset has been treated to protect confidentiality. Several methods have been applied to protect the confidentiality: removing variables from the original dataset, reducing detail in variables by recoding and top-coding and removing particular values of individual records at risk (local suppression) as well as randomization of the record order and record ids.

Removing variables

The released microdata set has only a selected number of variables contained in the initial survey. Not all variables could be released in this PUF without breaching confidentiality rules. Table 2 gives an overview of the removed variables.

Table 2 Overview of removed variables by module

Variable name	**Description**
Module 1	
m1siq1	Name of enumerator during first visit
m1siq2	Name of enumerator during second visit
m1siq4a	Does this school offer classes specifically for children with special needs/disabilities?
m1siq5a	School Name
m1siq5b	School Survey Code
m1siq6	School EMIS Code/Registration Number
m1siq7N	GPS Position N
m1siq7E	GPS Position E
m1saq0	Name of respondent
m1saq2	Respondents contact (Phone No.)
m1sdq9ac	For each term when did the school open and close?-End of term 1
m1sdq9bc	For each term when did the school open and close?-End of term 2
m1sdq9cc	For each term when did the school open and close?-End of term 3
m1sdq7_1e	Ending time of school day – First shift

m1sdq7_1e	Ending time of school day – Second shift
Module 2	
m2saq2	First and last names
m2sbq1	First and last names
Module 4	
m4siq1a	Enumerator name
m4siq3a	School Name
m4siq3b	School Code
m4siq4	School EMIS Code/Registration Number
m4siq6	Scheduled class time
m4siq8	Teacher name
Module 5	
m5siq1	Enumerator name
m5siq3	School Name
m5siq4	School Survey Code
m5siq8a	Teacher name
m5sa1q2	Pupil's first name
m5sa1q8a	Name of your English/Kiswahili teacher this year
m5sa1q9a	Name of your Math teacher this year
m5sa1q11a	Name of your English/Kiswahili teacher last year
m5sa1q12a	Name of your Math teacher last year

In several variables, the detail is reduced by recoding values or top-coding. The following table gives an overview of the variables that have been changed.

Table 3 Recodes of key variables

Variable name	**Description**	**Approach used**	**Before**	**After**
school level				
m1siq2a	Region	Group to Dar es Salaam and other	Names	Dar es Salaam; Other
m1siq4	Rural/urban	Group urban and semi-urban	1; 2, 3	1; 2
m1saq3	School ownership type	Public/private, other to missing	1; 2, 3	1; 2
m1saq6	Begin operating year	Group by decennia, bottom-code at 1960, top-code at 2000		1960, 70, 80, 90, 2000
m2sbq19	Number of classrooms	0, 1-9, 10-19, 20+	0; 1-9; 10-19; 20+	0; 5; 15; 20
teacher level				
m2saq4	Position in school	Group to head/teacher	1-4; 5-9	1;5
m2saq5	Contract status	Group “other”	1;2;3-6	1;2;3

Local suppression by variable

The objective of recoding is to achieve at least 2-anonimity in the data (remove sample uniques). This is achieved by suppressing certain values in the key variables. This process is done once for the variables at the school level and once for the variables at the teacher level. These vectors order the key variables according to their importance for the data user. Table 5 shows the number of suppression necessary to achieve 2-anonimity by variable after the recodings done in scenario 1 and scenario 2.

Table 4 Overview of number of suppressions per key variable

Variable name	**Description**	**Number of supressions**	**Total number of observations**
School level			
m1siq2a	Region	1	400
m1siq4	Rural/urban	0	400
m1saq3	School ownership type	0	400
m1saq4	School type	1	400
m1saq6	Begin operating year	11	400
m2sbq19	Number of classrooms	3	400
Teacher level			
schid	School id	12	6,964
m2saq3	Gender	0	6,964
m2saq4	Position in school	2	6,964
m2saq5	Contract status	0	6,964
m2saq6	Full / part time	0	6,964

Other measures

To guarantee confidentiality, the following measures are implemented in the R code:

- recode and randomize the district ids within the regions.
- randomize order and ID of schools within the regions. The initial order of the file would allow re-identification based on the order: two facilities that are close to one another in the file are also geographically close, since the file was ordered by lower level geographical variables. The mapping of the old to the new IDs is available.
- Randomize teacher IDs within schools. The mapping of the old to the new IDs is available.
- Randomize enumerator codes. The mapping of the old to the new IDs is available.
- Dates are recoded relative to the day of the survey (variables m1siq8, m1siq9, m1siq10d, m1siq10e, m4siq7, m5siq5, m6siq4)
- m1sbq8 (when was the last visit of the official government quality assurance officer or inspector?) is recoded to months past until survey date
- Start and end times are recoded to duration and AM/PM (variables m1siq11, m1siq12, m1siq13, m1siq14, m5sa1q14, m5sa1q15)
- m4sbq1 (number of pupils in the room) is rounded to the nearest 10 and top-coded at 50.

- m4sqd2 (number of pupils registered in class) is rounded to the nearest 10
- m5saq3 (age of pupils) is bottom-coded at 9 and top-coded at 13
- Number of toilets (m1scq3 and m1scq4) is top-coded at 10. The subcategories m1scq3a and m1scq4b are top-coded at 2 and m1scq3a is top-coded at 1.
- Recode absolute number to proportions. This is important when the aggregate is recoded. The proportions are computed from the original values (m2sbq20, m2sbq21, m2sbq22a, m2sbq22b, m2sbq22c from m2sbq19 (number of classrooms), m4sbq2, m4sbq3, m4sbq4, m4sbq4a, m4sbq4b, m4sbq5, m4sbq5a, m4sbq5b, m4sbq6, m4sbq6a, m4sbq6b, m4scq2, m4scq2a, m4scq2b, m4scq5, m4scq5a, m4scq5b, m4scq6, m4scq6a, m4scq6b, m4scq10, m4scq10a, m4scq10b, m4scq12 and m4saq5a – m4saq60a from m4sbq1 (number of pupils in the room), m4sdq3 from m4sdq2 (number of pupils registered in class)).
- Set all values in m4saqc3 and m4saqc4 to NA (count of proportion doesn't make sense)
- The variables m1bs9 (toilet facility) and m1sbq10 (water) are recoded to improved/not improved and clean/not clean, respectively.
- Recode the variables m2sbq9a, m2sbq9b, m2sbq9c and m2sbq9d (In which classes do you teach?) to lower and upper primary.
- Set values 5 and 6 in variables m2sb15a1-4 top NA (data error)
- Recode m1sdq9oa, m1sdq9bo and m1sdq9co to length of session
- Recode m1sdq7_1s and m1sdq7_2s to length of school day in hours
- Recode the values in m1sdq2a-c to proportions of the totals an suppress the totals
- Age of teachers (m2sbq14, m4sdq8, m6siq8) is recoded as follows: <25, 25-34, 35-44, 45-54, 55+
- Year of beginning teaching (m2sbq12, m4sdq11, m4sdq12) is recoded <1990, 1990-1999, 2000-2010, 2010+.
- Several variables are recoded according to the recodes of the key variables

Randomization

To further anonymize the data, the order of the records was randomized. Also teacher and school IDs were randomized. The randomized IDs still allow to match the files. Randomization allows using these variables for the evaluation of fixed effects.

5. Note on information loss

SDC methods lead to a loss of information or utility in the data. To check the validity of the data for research purposes, several indicators were computed from the original data and from treated data. The indicators values computed from the released dataset do not significantly differ from the values in the original dataset and hence the dataset is valid for research purposes.

6. Bibiliography

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., et al. (2012). *Statistical Disclosure Control.* Chichester, UK: John Wiley & Sons Ltd.

Templ, M., Meindl, B., Kowarik, A., & Chen, S. (2014, August 1). *Introduction to Statistical Disclosure Control (SDC).* Retrieved July 14, 2015 from http://www.ihsn.org/home/sites/default/files/resources/ihsn-working-paper-007-Oct27.pdf