# Nigeria 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 Nigeria 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 Nigeria education data set

The SDI Nigeria education dataset contains information on a sample of 760 schools from a total of about 8,500 schools in four regions (Anambra, Bauchi, Ekiti, Niger). 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 Nigeria. 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) |
|---|---|
| m1saq1 (position of respondent in facility) | 9 |
| m1scq9 (pupil toilet type) | 99 |
| m1scq10 (drinking water source) | 99 |
| m2saq5 (contract status) | 9 |
| m4scq5 (number of pupils) | -6 |

Other checks include reviewing outliers and dealing with facilities that may have refused to complete a part of the questionnaire. Furthermore, 16 schools without valid weights were removed from the released datasets.[3]

## 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 1 gives an overview of the removed variables.

[1] Available from https://www.r-project.org/.
[2] The package is available from CRAN (https://cran.r-project.org/).
[3] These are the schools with school IDs 21034, 21169, 21173, 22089, 22098, 31001, 31009, 31013, 32082, 32108, 32115, 32175, 42066, 42122, 42198 and 42199. Please confirm that these are the IDs from the original data set.

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 |
| m1siq5a | School Name |
| m1siq5b | School Survey Code |
| m1siq6 | School EMIS Code/Registration Number |
| m1siq7S | GPS Position S |
| m1siq7E | GPS Position E |
| m1saq0 | Name of respondent |
| m1saq2 | Respondents contact (Phone No.) |
| *Module 2* | |
| m2saq2 | First and last names |
| m2sbq1 | First and last names |
| 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 |
| *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 |
| *Module 6* | |
| m6siq4 | District code |
| m6siq5a | School name |
| m6siq5b | School code |

| |
|---|
| All variables with variable names ending in ab, bb, cb, db, eb, fb, gb, hb, ib, jb, kb |
| All variables containing characters (answers to test questions) |

### Reducing detail in variables by recoding and top-coding

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* | | | | |
| m1siq4 | Rural/urban | Group urban and semi-urban | 1, 3; 2 | 1; 2 |
| m1saq3 | School ownership type | Public/private, other to missing | 1; 2, 3 | 1; 2 |
| m1saq4 | School type | Group boarding and/or day school | 1,2; 3 | 1;2 |
| m1saq5 | School category | Group single-gender schools | 1,2;3 | 1;2 |
| m1saq6 | Begin operating year | Group by decennia and bottom-code at 1950 | | 1950, 60, 70, 80, 90, 2000, 2010 |
| 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

Values of certain variables for particular schools and teachers were deleted. Table 3 gives an overview of the number of suppressed values per variable.

Table 4 Overview of number of suppressions per key variable

| Variable name | Description | Number of supressions | Total number of observations |
|---|---|---|---|
| *School level* | | | |
| m1siq2a | Region | 2 | 744 |
| m1siq4 | Rural/urban | 0 | 744 |
| m1saq3 | School ownership type | 2 | 744 |
| m1saq4 | School type | 8 | 744 |
| m1saq5 | School category | 21 | 744 |
| m1saq6 | Begin operating year | 18 | 744 |
| m2sbq19 | Number of classrooms | 2 | 744 |
| *Teacher level* | | | |
| schid | School id | 16 | 7,673 |

| | | | |
|---|---|---|---|
| m2saq3 | Gender | 0 | 7,673 |
| m2saq4 | Position in school | 0 | 7,673 |
| m2saq5 | Contract status | 69 | 7,673 |
| m2saq6 | Full / part time | 0 | 7,673 |

## Other measures

To guarantee confidentiality, the following measures were implemented:

- 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, m1sbq8_a, m1sbq8_b, m1sbq8_c, m4siq7, m4siq10e, m4siq10f, m5siq5, m5siq9e, m5siq9f, m6siq4)
- Start and end times are recoded to duration and AM/PM (variables m1siq11, m1siq12, m1siq13, m1siq14)
- Recode the values in m1sdq2a-c to proportions of the totals an suppress the totals
- 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
- m5siq3 (age of pupils) is bottom-coded at 9 and top-coded at 13
- Recode absolute number to proportions. This is important when the aggregate is recoded. The proportions are computed from the original values (m2sbq20, m2sbq21 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)).
- The variables m1bs9 (toilet facility) and m1sbq10 (water) are recoded to improved/not improved and clean/not clean, respectively.
- Set all values in m4saqc3 and m4saqc4 to NA (count of proportion doesn't make sense)
- Recode the variables m2sbq9a, m2sbq9b, m2sbq9c and m2sbq9d (In which classes do you teach?) to lower and upper primary.
- Age of teachers (m2sbq12, m4sdq8, m6siq8) is recoded as follows: <25, 25-34, 35-44, 45-54, 55+
- Year of beginning teaching (m2sbq12, m4sdq11, m6siq13) and year obtained highest degree (m6siq16) is recoded <1990, 1990-1999, 2000-2010, 2010+.
- Several variables are recoded according to the recodes of the key variables

The variable and variable labels in the dataset have been adapted accordingly.

## 5. 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.

## 6. 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.

## 7. Bibliography

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