Data Cleaning Guide for PETS/QSDS Surveys

A Comprehensive Data Cleaning Guideline Applicable to Future PETS/QSDS Surveys

Contents

1. Introduction

This guide provides data quality analysts and researchers with a standard cleaning methodology to be used for work on future Public Expenditure Tracking Survey (PETS) and Quantitative Service Delivery Survey (QSDS) data sets. The objective of the data cleaning effort is to populate a Web-based PETS/QSDS data platform destined for use researchers, civil society and other stakeholders.

The current data cleaning effort focuses on the identification of fatal flaws in the questionnaire, as well as erroneous data entry and high-level internal coherence analysis. Wherever possible, data quality will be improved; however, it is essential that it at least be retained. Following the data cleaning process described here, the data should be both more credible and more user-friendly for analysis purposes. The extent to which additional data cleaning – e.g. more extensive data edits and imputation, narrower outlier analysis – should be applied is determined by the data's expected purpose of use; as such, it remains in the hands of final data users.

This guide provides data quality analysts with step-by-step instructions based on STATA 10 to verify the structural stability of new PETS/QSDS data sets, identify invalid entries and determine the data points that should be subjected to editing or imputation. The guide also describes the output of the data cleaning effort, the data quality report.

While most steps can be conducted on incomplete samples, the outlier analysis should only be performed on complete data sets to ensure coherent results.

2. Data Cleaning Procedure

Data cleaning is a three-step procedure:

a. Verification of structural stability: Questionnaires are complete, variables and responses are labeled accurately with regards to questionnaire labels, non-response codes are correctly entered, etc.

b. Identification of invalid entries: Out-of-range values, inconsistent responses across questions, outliers (high or low).

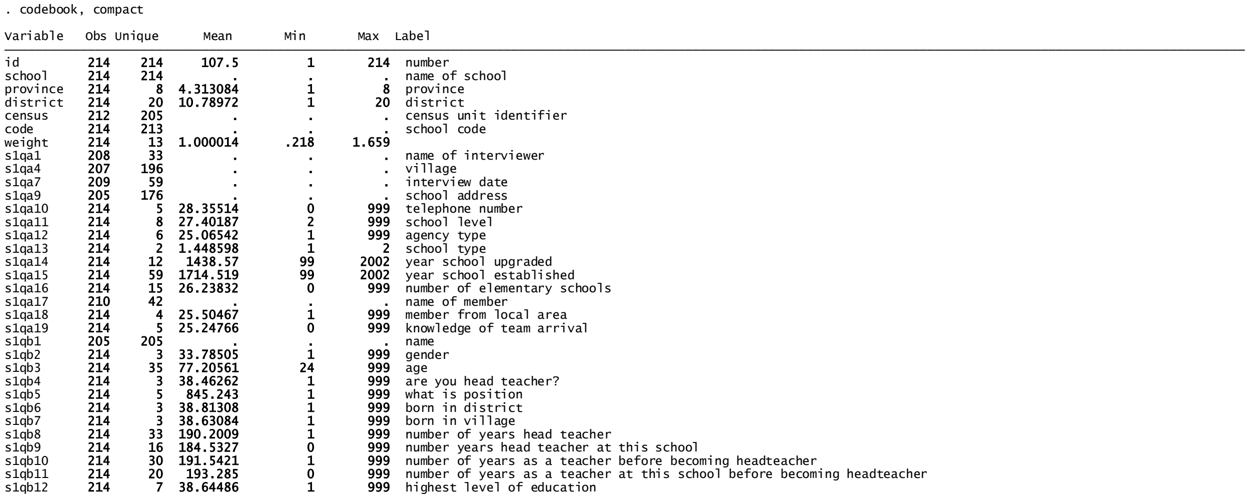
c. Editing and imputation.

a. Verification of Structural Stability

The verification of a data file's structural stability is a tedious process that requires systematic referral to the questionnaire in order to validate variables, values, labels, etc.

Step 1: Ensure all variables are present and labeled.

Use STATA's compact codebook (command *codebook, compact)*, which lists variable names and labels, as well as the number of total and unique observations, and variables' maxima and minima.



```
. codebook, compact

Variable    Obs Unique      Mean      Min      Max  Label
------------------------------------------------------------------------------------------------------------------------
id          214    214     107.5        1      214  number
school      214    214         .        .        .  name of school
province    214      8  4.313084        1        8  province
district    214     20  10.78972        1       20  district
census      212    205         .        .        .  census unit identifier
code        214    213         .        .        .  school code
weight      214     13  1.000014     .218    1.659
s1qa1       208     33         .        .        .  name of interviewer
s1qa4       207    196         .        .        .  village
s1qa7       209     59         .        .        .  interview date
s1qa9       205    176         .        .        .  school address
s1qa10      214      5  28.35514        0      999  telephone number
s1qa11      214      8  27.40187        2      999  school level
s1qa12      214      6  25.06542        1      999  agency type
s1qa13      214      2  1.448598        1        2  school type
s1qa14      214     12   1438.57       99     2002  year school upgraded
s1qa15      214     59  1714.519       99     2002  year school established
s1qa16      214     15  26.23832        0      999  number of elementary schools
s1qa17      210     42         .        .        .  name of member
s1qa18      214      4  25.50467        1      999  member from local area
s1qa19      214      5  25.24766        0      999  knowledge of team arrival
s1qb1       205    205         .        .        .  name
s1qb2       214      3  33.78505        1      999  gender
s1qb3       214     35  77.20561       24      999  age
s1qb4       214      3  38.46262        1      999  are you head teacher?
s1qb5       214      5   845.243        1      999  what is position
s1qb6       214      3  38.81308        1      999  born in district
s1qb7       214      3  38.63084        1      999  born in village
s1qb8       214     33  190.2009        1      999  number of years head teacher
s1qb9       214     16  184.5327        0      999  number years head teacher at this school
s1qb10      214     30  191.5421        1      999  number of years as a teacher before becoming headteacher
s1qb11      214     20   193.285        0      999  number of years as a teacher at this school before becoming headteacher
s1qb12      214      7  38.64486        1      999  highest level of education
```

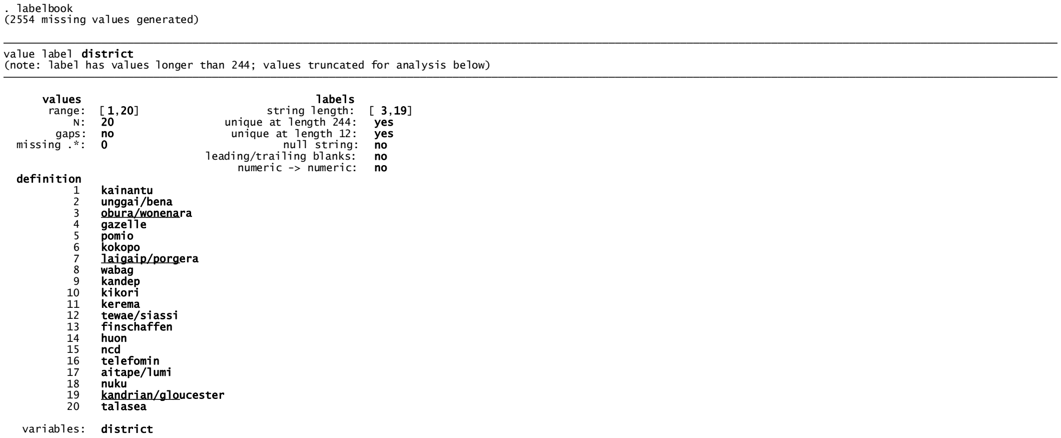
Figure 1: CODEBOOK, COMPACT Output

Using the *codebook, compact* output, verify that every variable listed in the data file – the variable names are in the first column – also exist in the questionnaire, and that no variables have been omitted. Any discrepancy between the data file and the questionnaire should appear in the data quality report.

While reviewing each variable, simultaneously confirm that each variable has been labeled, and that the label is relevant. If variables are missing labels, these can be added using STATA command *label variable* *varname “label”* where *varname* is the name of the variable listed in the first column and *label* is the relevant label. Labels can have up to 80 characters.

Step 2: Ensure response value labels are correct.

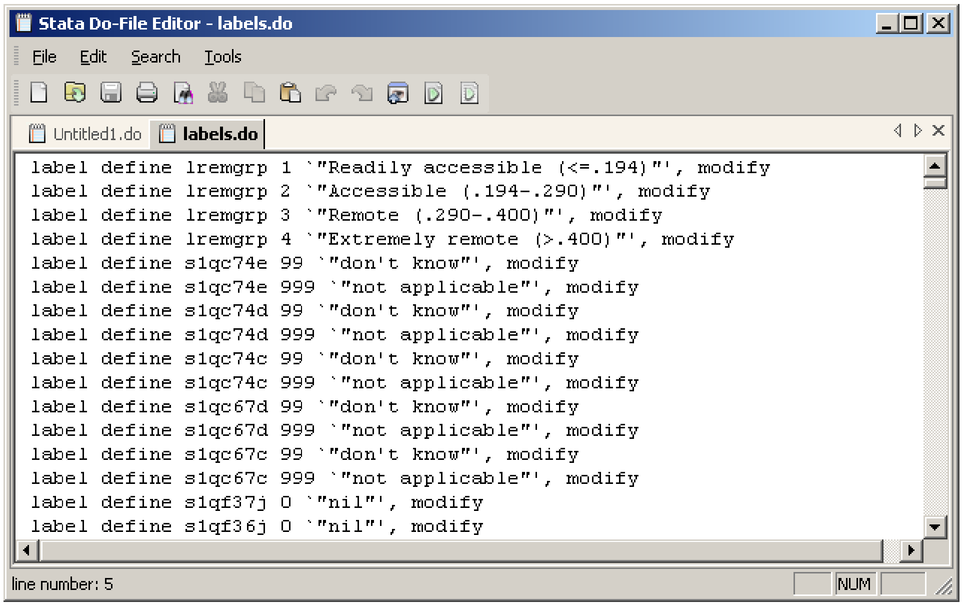
Compare variable labels (correct answers to categorical questions) in the questionnaire to those in the data set, to ensure the correct encoding of data. Existing value labels can be obtained using STATA command *labelbook*. The command lists the allowable labels under *definition*, as well as the variables the label applies to.



```
. labelbook
(2554 missing values generated)

-------------------------------------------------------------------------------
value label  district
(note: label has values longer than 244; values truncated for analysis below)
-------------------------------------------------------------------------------

      values                                    labels
       range:  [1,20]                     string length:  [ 3,19]
           N:  20                 unique at length 244:   yes
        gaps:  no                  unique at length 12:   yes
  missing .*:  0                           null string:   no
                              leading/trailing blanks:   no
                                   numeric -> numeric:   no

  definition
           1   kainantu
           2   unggai/bena
           3   obura/wonenara
           4   gazelle
           5   pomio
           6   kokopo
           7   laigaip/porgera
           8   wabag
           9   kandep
          10   kikori
          11   kerema
          12   tewae/siassi
          13   finschaffen
          14   huon
          15   ncd
          16   telefomin
          17   aitape/lumi
          18   nuku
          19   kandrian/gloucester
          20   talasea

   variables:  district
```

Figure 2: LABELBOOK Output

Alternatively, value labels can be obtained from the *Data -> Labels -> Label Values -> Save label values as do-file* menu option. The output, displayed below, can be opened in a text editor (STATA Do Editor, Notepad, Word) to facilitate cross-referencing with the questionnaire.



```
label define lremgrp 1 `"Readily accessible (<=.194)"', modify
label define lremgrp 2 `"Accessible (.194-.290)"', modify
label define lremgrp 3 `"Remote (.290-.400)"', modify
label define lremgrp 4 `"Extremely remote (>.400)"', modify
label define s1qc74e 99 `"don't know"', modify
label define s1qc74e 999 `"not applicable"', modify
label define s1qc74d 99 `"don't know"', modify
label define s1qc74d 999 `"not applicable"', modify
label define s1qc74c 99 `"don't know"', modify
label define s1qc74c 999 `"not applicable"', modify
label define s1qc67d 99 `"don't know"', modify
label define s1qc67d 999 `"not applicable"', modify
label define s1qc67c 99 `"don't know"', modify
label define s1qc67c 999 `"not applicable"', modify
label define s1qf37j 0 `"nil"', modify
label define s1qf36j 0 `"nil"', modify
```

Figure 3: Label Values in Text File

Any discrepancies between the questionnaire and the data set should be noted in the data quality report.

Step 3: Ensure each entry has a unique identifier

In many cases, the data will naturally contain a unique identifier (an establishment number, an employee number, a district number, etc). If this is the case, STATA function “codebook, compact” can help ensure each entry’s ID is unique, by comparing the *Obs* (Number of observations) and *Unique* (Number of unique entries) columns for that variable.

If a unique variable ID does not exist, the following line of STATA code can be used to create one:

(i) gen long id = _n;
(ii) lab var id "Entry ID";
(iii) compress id;

Line (i) creates a variable named “id” that can accommodate more than 2 billion entries. Line (ii) assigns label “Entry ID” to variable “id”. Finally, line (iii) reduces the size of variable “id” if, and only if, this leaves the underlying data unchanged.

As the unique identifier is usually used to link various databases together, its name should be entered in the data quality report.

Step 4: Recode missing values

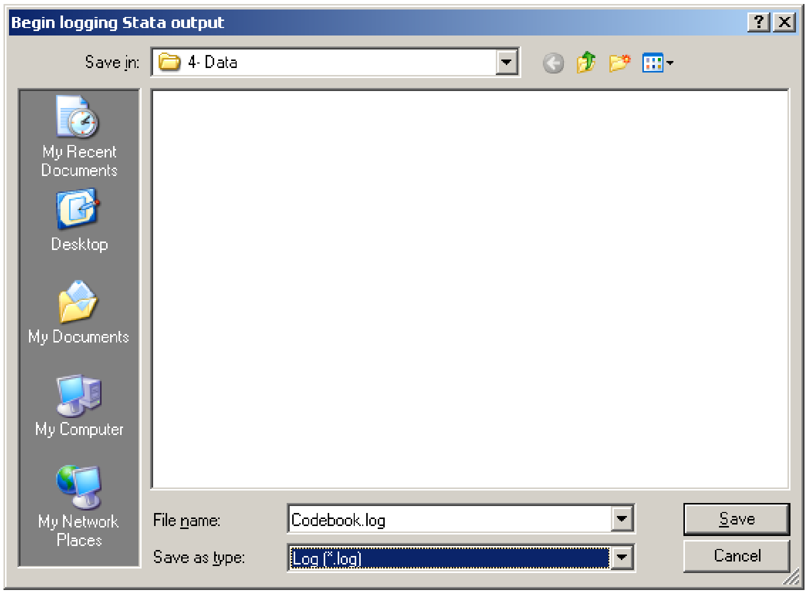
Missing values must be harmonized across all PETS/QSDS data sets, independently of missing value codes suggested by individual questionnaires. Following best practices, negative 3-digit integers should be used in order to ensure there is no confusion between missing values and valid data points. It is suggested that the following codes be used: DON’T KNOW (-666), NOT APPLICABLE (-777), LACK OF RECORDS (-888) and REFUSED TO ANSWER (-999).

Statistical packages such as STATA allow for differentiated missing value codes for analysis purposes; encoding is strongly facilitated by using differentiated codes for each of the missing values. In order to recode variables, use STATA command *mvdecode list_of_variables, mv(old_value=new missing value)*. The procedure can be applied to more than one variable at once by listing all variables, separated by a space, in *list_of_variables*, so long as they have the same old to new value conversion. Furthermore, more than one value can be replaced at once; *old_value = new_value* pairs are separated by a backslash (\).

Ex: *mvdecode var1 var2 var3, mv(99=-666 \ 999 = -777)* will replace values 99 with -666 and 999 with -777 for variables *var1*, *var2* and *var3*.

Step 5: Create a codebook and dictionary file for the dataset.

Create a codebook file that includes variable names and labels, type and length. The codebook can be created by first initiating STATA’s *log* function (menu: FILE -> LOG -> BEGIN), using function *codebook,* *compact* and stopping the LOG function (FILE -> LOG -> CLOSE). When beginning the log, you will be asked to enter a log file name. Ensure a representative name for the codebook is selected, and that the file type is *log*, not *SMCL*.



Save in: 4- Data

File name: Codebook.log

Save as type: Log (*.log)

Figure 4: Creating LOG File - The blue box must read LOG

Create a dictionary file that will allow data files to be imported into STATA. A dictionary file can be created using STATA command *outfile using dataset1.dct, dictionary* where *dataset1* is the name of the dataset.

The name of the dictionary file should be entered in the data quality report.

b. Identification of Invalid Entries

The identification of invalid entries is a two-part process. First, the questionnaire's internal coherence – skip patterns, sums – must be verified. Second, individual responses should be checked, first for legality, coherence and plausibility (outlier tests). To facilitate data analysis, it is recommended that missing values be recoded to STATA missing values using:

mvdecode _all, mv(-666 = .d \ -777 = .n \ -888 = .l \ -999 = .r).

The data cleaning do-file, which contains the list of error and coherence tests, will be appended to the data quality report.

i. Verifying Internal Coherence

Step 6: Identify skip patterns

Browse the questionnaire to identify questions that could be skipped due to previous answers. Every question that entails a possible skip should be noted, and a STATA test written to identify incoherent response patterns. For each skip pattern identified in the questionnaire, create an Error variable, incrementing the number that follows Error.

For example, if Question 1 is a yes (1) or no (2) question, and Question 2 begins with "If yes", then for any observation where the answer to Question 1 is 2, Question 2 should be skipped. In such a case, the error test would be:

Error01 = (Question1 == 2 & Question2 < .)

If the answer to question 1 is no (2) (important: note the two equal signs used to test equality) and question 2 is not a missing value (STATA stores missing values as the largest possible numbers, so any non-missing response is lower than a missing value), Error1 will take on the value of 1 (True).

If the subsequent skip is based on an alphabetic rather than numeric variable, the equivalent to a missing value is an empty string, with symbol *!=* representing *not equal to* and consecutive quotation marks *""* representing a missing or empty string:

Error02 = (Question1 == 2 & Question2 != "")

Step 7: Identify other internal coherence issues

Browse the questionnaire to identify implicit summations. These can include, among others, percentage sums that should total 100% or disaggregated totals (Total number of pupils, then disaggregation by gender). Error tests should be created to test the summation:

Error03 = (PercentMale + PercentFemale != 100)

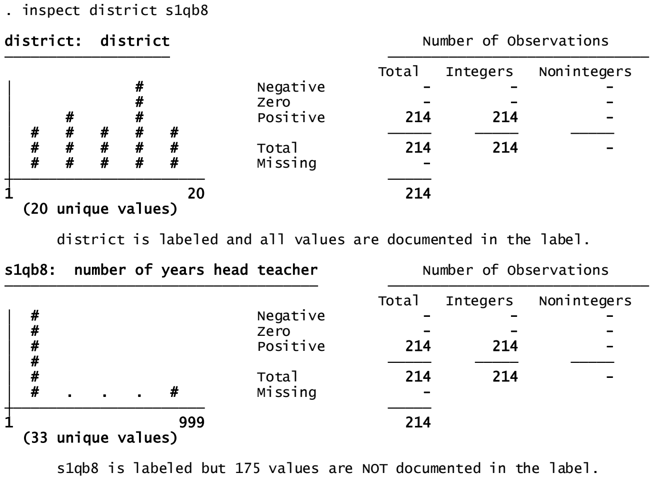
Error04 = (MalePupil + FemalePupil != TotalPupil)

ii. Verifying Individual Responses

Individual responses can be incorrect for one of three reasons: they can be illegal given the value labels assigned to a variable, can be inconsistent with other responses provided, or can be too far away from the mean response to have conceivably be drawn from the same random sample. In the third case, while the value may not be incorrect, it warrants further review, and may be dropped for statistical purposes.

Step 8: Identify illegal responses

Responses to categorical variables are illegal if they do not fall within the categories defined in the questionnaire. STATA command *Inspect* provides the necessary information to verify value labels. If a variable has a value label, STATA prints the following: "*varname is labeled and all values are documented in the label*". If there are illegal values, STATA publishes the following: *"varname is labeled but ## values are NOT documented in the label"*.



```
. inspect district s1qb8

district:  district                             Number of Observations
-------------------                         ------------------------------
                                           Total   Integers   Nonintegers
|              #             Negative           -         -          -
|              #             Zero               -         -          -
|      #       #             Positive         214       214          -
|  #   #   #   #   #                        -----     -----      -----
|  #   #   #   #   #         Total            214       214          -
|  #   #   #   #   #         Missing            -
+-----------------------                    -----
1                      20                     214
  (20 unique values)

      district is labeled and all values are documented in the label.

s1qb8:  number of years head teacher            Number of Observations
------------------------------------        ------------------------------
                                           Total   Integers   Nonintegers
|  #                         Negative           -         -          -
|  #                         Zero               -         -          -
|  #                         Positive         214       214          -
|  #                                        -----     -----      -----
|  #                         Total            214       214          -
|  #   .   .   .   #         Missing            -
+-----------------------                    -----
1                     999                     214
  (33 unique values)

      s1qb8 is labeled but 175 values are NOT documented in the label.
```

Figure 5: INSPECT Function Output

Note variables containing illegal values, and inspect them manually.

Step 9: Identify incoherent responses

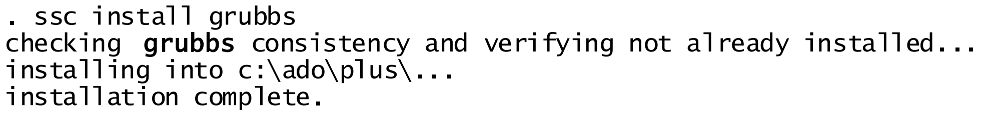
Responses can also be legal, but incoherent. For example, if consecutive questions ask i) how long a staff member has worked in a health center and ii) how long a staff member has *managed* a health center, it can be asserted that the answer to ii) should be less than or equal to the answer to i). Such errors should be identified, in a manner similar to the consistency errors in the previous section:

Coherence01 = (Question2 >= Question1)

Step 10: Identify outliers

Variables that are not bound by labels should be verified for the presence of outliers. While outliers are neither illegal nor incoherent values, they are values that do appear statistically unlikely to have been drawn from the same distribution as the rest of the sample.

A simple test to identify outliers is the Grubbs, or maximum normed residual, test. The Grubbs test is not installed by default on STATA 10; however, it can be installed by typing *ssc install grubbs*. STATA will confirm the successful installation of the command.



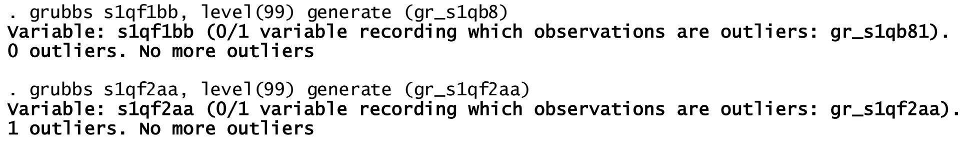
```
. ssc install grubbs
checking grubbs consistency and verifying not already installed...
installing into c:\ado\plus\...
installation complete.
```

Figure 6: GRUBBS Installation Output

In order to implement the Grubbs test, use the STATA command below. The command simultaneously checks each of the variables in *var1* to *varN* (separated by a space) and generates, for each variable, a new variable *grubbs_var1* to *grubbs_varN* that takes on the value of 1 if the observation is an outlier.

Given that the objective of the PETS/QSDS is to create a coherent database while keeping as many observations as possible, the confidence level is set extremely high, at 99.999% (4 standard deviations from the mean).

grubbs var1 var2 … varN, level(99.999)



```
. grubbs s1qf1bb, level(99) generate (gr_s1qb8)
Variable: s1qf1bb (0/1 variable recording which observations are outliers: gr_s1qb81).
0 outliers. No more outliers

. grubbs s1qf2aa, level(99) generate (gr_s1qf2aa)
Variable: s1qf2aa (0/1 variable recording which observations are outliers: gr_s1qf2aa).
1 outliers. No more outliers
```

Figure 7: GRUBBS Function Output

Where currency variables are collected, outliers can also be looked at on the basis of a normalized ratio (by pupil, by teacher, by patient, by doctor if relevant).

c. Editing and Imputation

In a context such as that of putting together a meta-database to be used by a broad range of researchers, the objective of data quality control is to provide data that is fully coherent, but more importantly, that is as close as possible to the information provided by the respondent. As such, the sanctity of the data is paramount. Data is to be edited only if it is deemed incorrect “beyond any reasonable doubt” and edited results are more likely to be accurate than the original data. Individual researchers can later determine if, for their research purposes, additional editing or imputation are justified.

Step 11: Identify errors to be reviewed

In order to avoid biasing responses through the application of a model, excessive imputation is strictly prohibited. If errors occur in 5% or more of responses, no imputation is allowed.

To determine which errors should be reviewed, first generate the codebook of error, coherence and outlier variables generated in the previous section: *codebook Error* Coherence* grubbs_*, compact*

The *Mean* column (see Figure 1 for Codebook output) determines the proportion of cases that exhibit the error. Any error, coherence or outlier variable with a mean in excess of 0.05 should be noted in the data quality report, but left unaltered.

Other error, coherence and outlier variables are to be reviewed.

Step 12: Reviewing errors and imputing responses

Data imputation should respect the principles elicited by Fellegi and Holt (1976), which assert that edits must be kept to a minimum number of fields required to pass consistency checks, as respondents are more likely to answer correctly than incorrectly.

Review the inconsistencies identified one by one, selecting the relevant variables for each error. For example, if Error01 calls upon variables Var01, Var02 and Var03, the following STATA command would generate the required subsample of data: *browse id var01 var02 var03 if Error01 == 1*. It is important to include variable *id* in order to be able to identify the incorrect observation.

For traceability purposes, data should be imputed using a do-file rather than the data browser. For example, if in the preceding example, it is determined that var01 should be set to 2 (NO) for observation 100, the following command would be entered: *replace Var01 = 2 if id == 100*. Note, again, the two equal (=) signs between *id* and the observation number.

Apparent inconsistencies between interlinked questions should first be considered by looking at the wording of specific questions. If the inconsistency can be explained by poor or inexact question wording, no editing should take place. Rather, the justification should be entered in the data quality report.

If question wording cannot explain inconsistencies, consider the response a respondent is most likely to have answered correctly. More cognitively difficult questions are more likely to be accurate, given the need for reflection on the respondent's behalf. Hence, if a respondent claims not having paid for security, but in an inter-linked question, claims having paid 8,500 dollars to security agents, the second data point will be assumed correct, and the first imputed.

With regards to the specific errors identified in the previous section, many internal coherence errors can be attributed not to truly incorrect responses, but rather to encoding errors, using 0s to represent missing values rather than the appropriate missing codes. If this error is identified, the 0 values should be replaced by missing values.

Finally, with regards to outliers, values that are outside a confidence interval of 99.999% should be looked at very closely. If the value cannot be explained, it may be replaced by a missing value.

At all times, edits introduced to the data should be noted in the code used to perform data quality control. Future data analysts must be able to reproduce every step from the initial to the publicly available dataset.

3. How to use PETS/QSDS data?

a. Data contents by survey

In each country, researchers can expect to find the following files:

- Survey tools (questionnaires, interview manuals);
- Files containing the raw data as collected and entered in the field;
- Files containing clean data following the PETS/QSDS Data Cleaning Guideline.
- Files used to convert the data sets from their raw to their clean versions.
- A data quality report that discusses, at a high level, the changes applied to the data as well as remaining issues in the data set.
- A data dictionary or codebook describing the variables contained in each data set.

Data files are provided in STATA 10 ©, but can be converted for use in other data analysis software (SPSS, SAS, Excel) using STAT/TRANSFER©.

b. Differences between collected and presented data sets

Each country set includes two data files. The first file, the "raw" data file, presents the data as collected and entered by the survey teams. While field teams do conduct very high-level coherence tests with regards to responses collected, the data contained therein has generally not been thoroughly checked for internal coherence across questions, variable outliers and other such involved data cleaning procedures.

The second file, the "final" data file, has been reviewed in order to ensure consistency both within and across single observations. While the sanctity of data is paramount, such that no changes are made if it cannot be asserted that the edited value is closer to the "true" value than the previous entry, data edits are introduced into the final data set. The list of edits applied are listed in the available Stata 10 © do-file associated with each data set. Furthermore, each do-file includes other tests that were applied to the data set. In addition, basic statistical analysis is applied to variables in order to identify potential statistical outliers. Outlier values that cannot be explained are replaced by missing values in the "final" data set; these changes are reported both in the do-file and in the Data Quality Report.

Finally, independently of the values presented in the questionnaires, missing values are replaced across all "final" data sets to ensure consistency across countries. Following industry best practices, negative 3-digit integers are used in order to ensure there is no confusion between missing values and valid data points. Across all data sets, the standard error codes are therefore:

- DON’T KNOW (-666);
- NOT APPLICABLE (-777);
- LACK OF RECORDS (-888); and
- REFUSED TO ANSWER (-999).

c. Using the available data files

This guide provides researchers with a high-level understanding of the PETS/QSDS surveys available across countries. However, each country survey is unique in its structure, such that additional information must be provided to make the data sets useable.

In order to assist researchers, the PETS/QSDS data platform provides a user guide that presents the data files available (e.g. Facilities survey, staff members survey, government official surveys, parent surveys, etc.) for each country survey. Furthermore, the country-specific user guide provides researchers with the instructions on how to merge the various data files available, in order to obtain a PETS/QSDS database that incorporates all relevant economic agents in a given country.

4. Annex: Data quality report

DATA QUALITY REPORT

1. DATA IDENTIFICATION	
Data File:	
Dictionary:	
Data Cleaning Do-File:	

2. STRUCTURAL STABILITY			
	Yes		No
2.1 Do all variables in the questionnaire match those in the data set?			
2.2 Do all value labels in the questionnaire match those in the data set?			
2.3 What is the name of the dataset's unique identifier?			

If 2.1 or 2.2 is false, please append additional sheets describing the discrepancies.

3. EDITING AND IMPUTATION			
	Yes		No
3.1 Did any variable exhibit errors in more than 5% of observations?			
3.2 Was any editing or imputation applied to the data set *due to illegal or incoherent responses*?			
3.3 Were any outliers rejected?			

If either 3.1 or 3.3 is true, please append additional sheets describing the errors identified.3.2 should appear in the attached data cleaning do-file.

Examples of additional sheets (these may be multiple pages long)

2.1 Variable mismatches

- Variables Var2 – Var5 in the questionnaire do not appear in the data set.
- Variables Var6 – Var9 in the questionnaire are equivalent to variables Var2 – Var5 in the data set.

2.2 Value label mismatches

- Variable Var6 in the dataset did not have value labels while the equivalent variable in the questionnaire did. The dataset variable was encoded according to the questionnaire categories.
- Variable Var7 in the dataset has value labels that do not appear in the questionnaire. The categories are as follows:
  1: 0-5 years
  2: 5-10 years
  3: 11-15 years
  4: more than 15 years

3.1 Variables with more errors in more than 5% of observations

- Variables Var8 and Var10 displayed errors in 7% and 12% of cases respectively; these variables were not corrected.

3.3 Outliers

- An outlier was rejected for observation 19, variable 4. It was converted to a missing value.